Sources of error

R.W. Oldford

UNIVERSITY OF WATERLOO

Population attributes:

Interest lies in assessing and/or discovering interesting **attributes** $a(\mathcal{P})$ of some population $\mathcal{P}$ of units $u \in \mathcal{P}$.

Population attributes:

Interest lies in assessing and/or discovering interesting **attributes** $a(\mathcal{P})$ of some population $\mathcal{P}$ of units $u \in \mathcal{P}$.

- units u are unique and distinct from one another
- often have many **variates** $x_1(u)$, $x_2(u)$, $\ldots$ associated with each unit, possibly
  - of different types (and scales)

Population attributes:

Interest lies in assessing and/or discovering interesting **attributes** $a(\mathcal{P})$ of some population $\mathcal{P}$ of units $u \in \mathcal{P}$.

- units u are unique and distinct from one another
- often have many **variates** $x_1(u)$, $x_2(u)$, ... associated with each unit, possibly
  - of different types (and scales)
  - of differing interpretability (e.g. physical measurements, summary calculations over different variates)

Population attributes:

Interest lies in assessing and/or discovering interesting **attributes** $a(\mathcal{P})$ of some population $\mathcal{P}$ of units $u \in \mathcal{P}$.

- units u are unique and distinct from one another
- often have many **variates** $x_1(u)$, $x_2(u)$, $\ldots$ associated with each unit, possibly
  - of different types (and scales)
  - of differing interpretability (e.g. physical measurements, summary calculations over different variates)
- a population attributes is any well defined summary of $\mathcal{P}$ and so could be
  - numerical

Population attributes:

Interest lies in assessing and/or discovering interesting **attributes** $a(\mathcal{P})$ of some population $\mathcal{P}$ of units $u \in \mathcal{P}$.

- units u are unique and distinct from one another
- often have many **variates** $x_1(u)$, $x_2(u)$, ... associated with each unit, possibly
  - of different types (and scales)
  - of differing interpretability (e.g. physical measurements, summary calculations over different variates)
- a population attributes is any well defined summary of $\mathcal{P}$ and so could be
  - numerical
  - graphical

Population attributes:

Interest lies in assessing and/or discovering interesting **attributes** $a(\mathcal{P})$ of some population $\mathcal{P}$ of units $u \in \mathcal{P}$.

- units u are unique and distinct from one another
- often have many **variates** $x_1(u)$, $x_2(u)$, ... associated with each unit, possibly
  - of different types (and scales)
  - of differing interpretability (e.g. physical measurements, summary calculations over different variates)
- a population attributes is any well defined summary of $\mathcal{P}$ and so could be
  - numerical
  - graphical
  - mathematical/algorithmic (e.g. a fitted model/function)
  - multidimensional

Population attributes:

Interest lies in assessing and/or discovering interesting **attributes** $a(\mathcal{P})$ of some population $\mathcal{P}$ of units $u \in \mathcal{P}$.

- units u are unique and distinct from one another
- often have many **variates** $x_1(u)$, $x_2(u)$, ... associated with each unit, possibly
  - of different types (and scales)
  - of differing interpretability (e.g. physical measurements, summary calculations over different variates)
- a population attributes is any well defined summary of $\mathcal{P}$ and so could be
  - numerical
  - graphical
  - mathematical/algorithmic (e.g. a fitted model/function)
  - multidimensional
- have many attributes $a_1(\mathcal{P})$, $a_2(\mathcal{P})$, ... each summarizing some different aspect of the population $\mathcal{P}$

Population attributes:

Interest lies in assessing and/or discovering interesting **attributes** $a(\mathcal{P})$ of some population $\mathcal{P}$ of units $u \in \mathcal{P}$.

- units u are unique and distinct from one another
- often have many **variates** $x_1(u)$, $x_2(u)$, ... associated with each unit, possibly
  - of different types (and scales)
  - of differing interpretability (e.g. physical measurements, summary calculations over different variates)
- a population attributes is any well defined summary of $\mathcal{P}$ and so could be
  - numerical
  - graphical
  - mathematical/algorithmic (e.g. a fitted model/function)
  - multidimensional
- have many attributes $a_1(\mathcal{P})$, $a_2(\mathcal{P})$, ... each summarizing some different aspect of the population $\mathcal{P}$

UNIVERSITY OF WATERLOO

Population attributes:

Each attribute is

- a function of the population $\mathcal{P}$

Population attributes:

Each attribute is

- a function of the population $\mathcal{P}$ and
- hence of any or all variates $x_1(u)$, $x_2(u)$, ...

Population attributes:

Each attribute is

- a function of the population $\mathcal{P}$ and
- hence of any or all variates $x_1(u)$, $x_2(u)$, ... and
- of any subset of units $u \in \mathcal{P}$ (e.g. as determined by values of some of the variates).

Population attributes:

Each attribute is

- a function of the population $\mathcal{P}$ and
- hence of any or all variates $x_1(u)$, $x_2(u)$, $\ldots$ and
- of any subset of units $u \in \mathcal{P}$ (e.g. as determined by values of some of the variates).

The quality of an attribute therefore depends upon the quality of any and all of these constituents.

Population attributes:

Each attribute is

- a function of the population $\mathcal{P}$ and
- hence of any or all variates $x_1(u)$, $x_2(u)$, $\ldots$ and
- of any subset of units $u \in \mathcal{P}$ (e.g. as determined by values of some of the variates).

The quality of an attribute therefore depends upon the quality of any and all of these constituents.

We need to consider what general sources might contibute to error (besides calculational/floating point errors).

Example: Surgery or radiation?

Suppose, we are interested in the proportion of people who would choose surgery over radiation when presented with the following scenario:

Example: Surgery or radiation?

Suppose, we are interested in the proportion of people who would choose surgery over radiation when presented with the following scenario:

> *"In decisions about patient care, both the physician and the patient will participate in determining the care and treatment which the patient will receive.*
>
> *Imagine the following hypothetical medical situation where you, the patient, having been diagnosed with a form of cancer are trying to make a choice between two different treatments available. The treatments are (a) Surgery and (b) Radiation. The decision as to which treatment you will take is entirely yours.*
>
> *To help you make an informed treatment, the physician presents you with the following information based on previous medical studies: "*

Which would then be followed by relevant numerical information on historical outcomes from patients who had surgery and from those who had radiation.

Example: Surgery or radiation?

Suppose, we are interested in the proportion of people who would choose surgery over radiation when presented with the following scenario:

> *"In decisions about patient care, both the physician and the patient will participate in determining the care and treatment which the patient will receive.*
>
> *Imagine the following hypothetical medical situation where you, the patient, having been diagnosed with a form of cancer are trying to make a choice between two different treatments available. The treatments are (a) Surgery and (b) Radiation. The decision as to which treatment you will take is entirely yours.*
>
> *To help you make an informed treatment, the physician presents you with the following information based on previous medical studies: "*

Which would then be followed by relevant numerical information on historical outcomes from patients who had surgery and from those who had radiation.

Questions:

- What is the population $\mathcal{P}$? What are its units?

Example: Surgery or radiation?

Suppose, we are interested in the proportion of people who would choose surgery over radiation when presented with the following scenario:

> *"In decisions about patient care, both the physician and the patient will participate in determining the care and treatment which the patient will receive.*
>
> *Imagine the following hypothetical medical situation where you, the patient, having been diagnosed with a form of cancer are trying to make a choice between two different treatments available. The treatments are (a) Surgery and (b) Radiation. The decision as to which treatment you will take is entirely yours.*
>
> *To help you make an informed treatment, the physician presents you with the following information based on previous medical studies: "*

Which would then be followed by relevant numerical information on historical outcomes from patients who had surgery and from those who had radiation.

Questions:

- What is the population $\mathcal{P}$? What are its units?
- How about variate(s)? What is the kind of variate(s)?

Example: Surgery or radiation?

Suppose, we are interested in the proportion of people who would choose surgery over radiation when presented with the following scenario:

> *"In decisions about patient care, both the physician and the patient will participate in determining the care and treatment which the patient will receive.*
>
> *Imagine the following hypothetical medical situation where you, the patient, having been diagnosed with a form of cancer are trying to make a choice between two different treatments available. The treatments are (a) Surgery and (b) Radiation. The decision as to which treatment you will take is entirely yours.*
>
> *To help you make an informed treatment, the physician presents you with the following information based on previous medical studies: "*

Which would then be followed by relevant numerical information on historical outcomes from patients who had surgery and from those who had radiation.

Questions:

- What is the population $\mathcal{P}$? What are its units?
- How about variate(s)? What is the kind of variate(s)?
- What population attribute is of interest?

Example: Surgery or radiation?

Suppose, we are interested in the proportion of people who would choose surgery over radiation when presented with the following scenario:

> *"In decisions about patient care, both the physician and the patient will participate in determining the care and treatment which the patient will receive.*
>
> *Imagine the following hypothetical medical situation where you, the patient, having been diagnosed with a form of cancer are trying to make a choice between two different treatments available. The treatments are (a) Surgery and (b) Radiation. The decision as to which treatment you will take is entirely yours.*
>
> *To help you make an informed treatment, the physician presents you with the following information based on previous medical studies: "*

Which would then be followed by relevant numerical information on historical outcomes from patients who had surgery and from those who had radiation.

Questions:

- What is the population $\mathcal{P}$? What are its units?
- How about variate(s)? What is the kind of variate(s)?
- What population attribute is of interest?
- What role is played by the question asked?

Example: Surgery or radiation?

A class of graduate students were split into four groups, each group receiving a slightly different presentation of the historical data.

All four groups had the same preamble about the question, just different "information based on previous medical studies".

Example: Surgery or radiation?

A class of graduate students were split into four groups, each group receiving a slightly different presentation of the historical data.

All four groups had the same preamble about the question, just different "information based on previous medical studies".

Groups 1 and 2:

- had the information shown as diagrams, one related to surgery outcomes, one related to radiation outcomes
- had slightly different descriptions attached to each diagram

Example: Surgery or radiation?

A class of graduate students were split into four groups, each group receiving a slightly different presentation of the historical data.

All four groups had the same preamble about the question, just different "information based on previous medical studies".

Groups 1 and 2:

- had the information shown as diagrams, one related to surgery outcomes, one related to radiation outcomes
- had slightly different descriptions attached to each diagram

Groups 3 and 4:

- had the information given as numbers, one set related to surgery outcomes, the other related to radiation outcomes
- had slightly different descriptions attached to the numbers

Example: Surgery or radiation?

A class of graduate students were split into four groups, each group receiving a slightly different presentation of the historical data.

All four groups had the same preamble about the question, just different "information based on previous medical studies".

Groups 1 and 2:

- had the information shown as diagrams, one related to surgery outcomes, one related to radiation outcomes
- had slightly different descriptions attached to each diagram

Groups 3 and 4:

- had the information given as numbers, one set related to surgery outcomes, the other related to radiation outcomes
- had slightly different descriptions attached to the numbers

In all cases, the historical information presented was **identical**.

Example: Surgery or radiation?

A class of graduate students were split into four groups, each group receiving a slightly different presentation of the historical data.

All four groups had the same preamble about the question, just different "information based on previous medical studies".

Groups 1 and 2:

- had the information shown as diagrams, one related to surgery outcomes, one related to radiation outcomes
- had slightly different descriptions attached to each diagram

Groups 3 and 4:

- had the information given as numbers, one set related to surgery outcomes, the other related to radiation outcomes
- had slightly different descriptions attached to the numbers

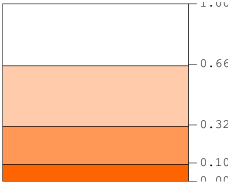
In all cases, the historical information presented was **identical**.

After the historical information was presented, each group was instructed:

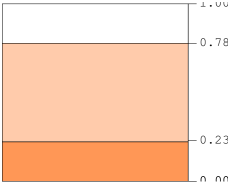
> *Based on this information, you must choose one of the two treatments. Circle one of the following as your answer:*
>
> ***(a) Surgery*** ***(b) Radiation***

Surgery or radiation: Groups 1 and 2 pictures presented.

In each diagram below the area of the horizontal strip is the probability of the outcome which labels the strip.



(a) Surgery



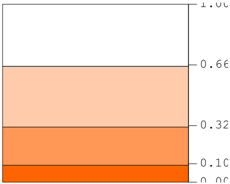
(b) Radiation

200 patients diagnosed with cancer

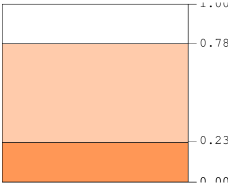
100 receive (a) surgery, 100 (b) radiation treatment.

Surgery or radiation: Group 1 was told

In each diagram below the area of the horizontal strip is the probability of the outcome which labels the strip.



(a) Surgery

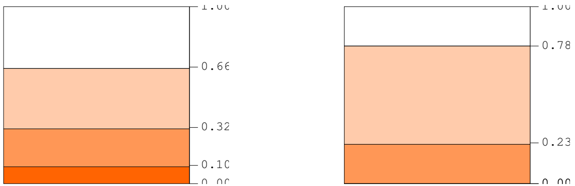


(b) Radiation

Figure 1: 200 patients diagnosed with cancer – 100 receive (a) surgery, 100 (b) radiation treatment.
From bottom to top the categories are y_1= "Does not survive treatment", y_2 = "Survives treatment, but only to one year, y_3= "Survives more than one but fewer than five years" and y_4 = "Survives at least 5 years".
The area (or equivalently the height) of each shaded rectangle matches the proportion of the 100 which are in that category.
The shading matches the category across the two figures and for radiation the bottom most category, y_1, is absent because all survive radiation treatment.

Surgery or radiation: Group 2 was told

In each diagram below the area of the horizontal strip is the probability of the outcome which labels the strip.



(a) Surgery (b) Radiation

Figure 2: 200 patients diagnosed with cancer – 100 receive (a) surgery, 100 (b) radiation treatment.
From bottom to top the categories are y_1= "Die during treatment", y_2 = "Die by the end of the first year, y_3= "Die by the end of five years" and y_4 = "Survives at least 5 years".
The area (or equivalently the height) of each shaded rectangle matches the proportion of the 100 which are in that category.
The shading matches the category across the two figures and for radiation the bottom most category, y_1, is absent because no one died during radiation treatment.

Surgery or radiation: Groups 3 and 4

Groups 3 and 4 were presented the historical information as text with numbers.

Group 3:

> ***(a) Surgery****: Of 100 people having surgery 90 live through the post-operative period, 68 are alive at the end of the first year, and 34 are alive at the end of five years.*
>
> ***(b) Radiation therapy****: Of 100 people having radiation therapy, all live through the treatment, 77 are alive at the end of one year, and 22 are alive at the end of five years.*

Surgery or radiation: Groups 3 and 4

Groups 3 and 4 were presented the historical information as text with numbers.

Group 3:

> ***(a) Surgery****: Of 100 people having surgery 90 live through the post-operative period, 68 are alive at the end of the first year, and 34 are alive at the end of five years.*
>
> ***(b) Radiation therapy****: Of 100 people having radiation therapy, all live through the treatment, 77 are alive at the end of one year, and 22 are alive at the end of five years.*

Group 4:

> ***Surgery****: Of 100 people having surgery 10 die during surgery or the post-operative period, 32 die by the end of the first year, and 66 die by the end of five years.*
>
> ***Radiation therapy****: Of 100 people having radiation therapy, none die during treatment, 23 die by the end of one year, and 78 die by the end of five years.*

Surgery or radiation: results

The objective was to determine the proportion p of people who would choose surgery.

Surgery or radiation: results

The objective was to determine the proportion p of people who would choose surgery.

	Surgery	Radiation	p

Surgery or radiation: results

The objective was to determine the proportion p of people who would choose surgery.

	Surgery	Radiation	p
Group 1	6	4	0.6

Surgery or radiation: results

The objective was to determine the proportion p of people who would choose surgery.

	Surgery	Radiation	p
Group 1	6	4	0.6
Group 2	6	4	0.6

Surgery or radiation: results

The objective was to determine the proportion p of people who would choose surgery.

	Surgery	Radiation	p
Group 1	6	4	0.6
Group 2	6	4	0.6
Group 3	6	4	0.6

Surgery or radiation: results

The objective was to determine the proportion p of people who would choose surgery.

	Surgery	Radiation	p
Group 1	6	4	0.6
Group 2	6	4	0.6
Group 3	6	4	0.6
Group 4	1	9	0.1

Surgery or radiation: results

The objective was to determine the proportion p of people who would choose surgery.

	Surgery	Radiation	p
Group 1	6	4	0.6
Group 2	6	4	0.6
Group 3	6	4	0.6
Group 4	1	9	0.1

There appear to be two very different values for the population attribute.

Surgery or radiation: results

The objective was to determine the proportion p of people who would choose surgery.

	Surgery	Radiation	p
Group 1	6	4	0.6
Group 2	6	4	0.6
Group 3	6	4	0.6
Group 4	1	9	0.1

There appear to be two very different values for the population attribute.

- What could have produced these differences?

Surgery or radiation: results

The objective was to determine the proportion p of people who would choose surgery.

	Surgery	Radiation	p
Group 1	6	4	0.6
Group 2	6	4	0.6
Group 3	6	4	0.6
Group 4	1	9	0.1

There appear to be two very different values for the population attribute.

- What could have produced these differences?

Giant redwoods: How high is the tallest California redwood?

Redwood trees (*sequoia sempervirens*) are an exceptionally tall tree that grows on the west coast of North America.

Giant redwoods: How high is the tallest California redwood?

Redwood trees (*sequoia sempervirens*) are an exceptionally tall tree that grows on the west coast of North America.

The following attributes are of interest:

1. the proportion of people who think the tallest redwood is higher than 50 metres

Giant redwoods: How high is the tallest California redwood?

Redwood trees (*sequoia sempervirens*) are an exceptionally tall tree that grows on the west coast of North America.

The following attributes are of interest:

1. the proportion of people who think the tallest redwood is higher than 50 metres
2. the proportion of people who think the tallest redwood is higher than 100 metres

Giant redwoods: How high is the tallest California redwood?

Redwood trees (*sequoia sempervirens*) are an exceptionally tall tree that grows on the west coast of North America.

The following attributes are of interest:

1. the proportion of people who think the tallest redwood is higher than 50 metres
2. the proportion of people who think the tallest redwood is higher than 100 metres
3. the average height that people think the tallest redwood cpuld be, in metres.

UNIVERSITY OF WATERLOO

Giant redwoods: How high is the tallest California redwood?

Redwood trees (*sequoia sempervirens*) are an exceptionally tall tree that grows on the west coast of North America.

The following attributes are of interest:

1. the proportion of people who think the tallest redwood is higher than 50 metres
2. the proportion of people who think the tallest redwood is higher than 100 metres
3. the average height that people think the tallest redwood cpuld be, in metres.

Questions:

- what is a population unit here?
- what is the population of interest?

Giant redwoods: How high is the tallest California redwood?

To get values for these population attributes, a class of graduate students were given the following:

Giant redwoods: How high is the tallest California redwood?

To get values for these population attributes, a class of graduate students were given the following:

1. Is the tallest California Redwood tree (Sequoia sempervirens) higher or lower than **A** metres tall?
   Circle one:

   Less than **A** metres   **MORE** than **A** metres.

2. Write down your best guess (in metres) of the tallest California Redwood tree:

Giant redwoods: How high is the tallest California redwood?

To get values for these population attributes, a class of graduate students were given the following:

1. Is the tallest California Redwood tree (Sequoia sempervirens) higher or lower than **A** metres tall?
   Circle one:

   Less than **A** metres | **MORE** than **A** metres.

2. Write down your best guess (in metres) of the tallest California Redwood tree:

The students were divided into two groups.

Giant redwoods: How high is the tallest California redwood?

To get values for these population attributes, a class of graduate students were given the following:

1. Is the tallest California Redwood tree (Sequoia sempervirens) higher or lower than **A** metres tall?
   Circle one:

   Less than **A** metres   **MORE** than **A** metres.

2. Write down your best guess (in metres) of the tallest California Redwood tree:

The students were divided into two groups.

For one group, **A** was replaced by **100**;

Giant redwoods: How high is the tallest California redwood?

To get values for these population attributes, a class of graduate students were given the following:

1. Is the tallest California Redwood tree (Sequoia sempervirens) higher or lower than **A** metres tall?
   Circle one:

   Less than **A** metres　　　　　　　　**MORE** than **A** metres.

2. Write down your best guess (in metres) of the tallest California Redwood tree:

The students were divided into two groups.

For one group, **A** was replaced by **100**; for the other, **A** was replaced by **50**.

Giant redwoods: Results

Data:

```
redwoods <- read.csv(path_concat(dataDirectory, "redwood.csv"))
# Last two rows
tail(redwoods, n = 2)
```

```
##     A more guess
## 37 50   no    35
## 38 50  yes   100
```

```
# Number A = 50
A_50 <- redwoods$A == 50
sum(A_50)
```

```
## [1] 19
```

```
# Number A = 100
A_100 <- redwoods$A == 100
sum(A_100)
```

```
## [1] 19
```

Giant redwoods: Results

Data:

```
redwoods <- read.csv(path_concat(dataDirectory, "redwood.csv"))
# Last two rows
tail(redwoods, n = 2)
```

```
##      A more guess
## 37 50   no    35
## 38 50  yes   100
```

```
# Number A = 50
A_50 <- redwoods$A == 50
sum(A_50)
```

```
## [1] 19
```

```
# Number A = 100
A_100 <- redwoods$A == 100
sum(A_100)
```

```
## [1] 19
```

Proportions:

Giant redwoods: Results

Data:

```
redwoods <- read.csv(path_concat(dataDirectory, "redwood.csv"))
# Last two rows
tail(redwoods, n = 2)
```

```
##      A more guess
## 37 50   no    35
## 38 50  yes   100
```

```
# Number A = 50
A_50 <- redwoods$A == 50
sum(A_50)
```

```
## [1] 19
```

```
# Number A = 100
A_100 <- redwoods$A == 100
sum(A_100)
```

```
## [1] 19
```

Proportions:

```
said_yes <- redwoods$more == "yes"
# Proportion think tallest is greater than 50 metres
round(sum(A_50 & said_yes)/sum(A_50), 2)
```

```
## [1] 0.84
```

```
# Proportion think tallest is greater than 100 metres
round(sum(A_100 & said_yes)/sum(A_100), 2)
```

```
## [1] 0.84
```

Giant redwoods: Results

Average tallest heights:

```
mean(redwoods$guess)
```

```
## [1] 125.9474
```

Giant redwoods: Results

Average tallest heights:

```r
mean(redwoods$guess)
```

```
## [1] 125.9474
```

But what about for each group?

Giant redwoods: Results

Average tallest heights:

```
mean(redwoods$guess)
```

```
## [1] 125.9474
```

But what about for each group?

```
mean(redwoods$guess[A_50])
```

```
## [1] 92.52632
```

```
mean(redwoods$guess[A_100])
```

```
## [1] 159.3684
```

Giant redwoods: Results

Average tallest heights:

```
mean(redwoods$guess)
```

```
## [1] 125.9474
```

But what about for each group?

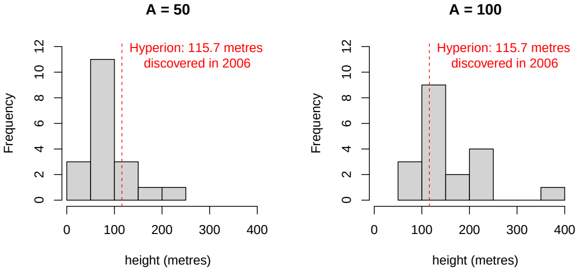
```
mean(redwoods$guess[A_50])
```

```
## [1] 92.52632
```

```
mean(redwoods$guess[A_100])
```

```
## [1] 159.3684
```

Histogram of tallest heights:



Giant redwoods: Results

Average tallest heights:

```
mean(redwoods$guess)
```

```
## [1] 125.9474
```

But what about for each group?

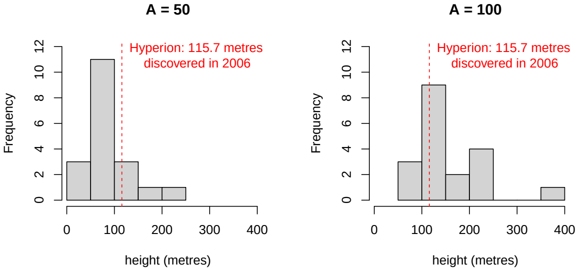
```
mean(redwoods$guess[A_50])
```

```
## [1] 92.52632
```

```
mean(redwoods$guess[A_100])
```

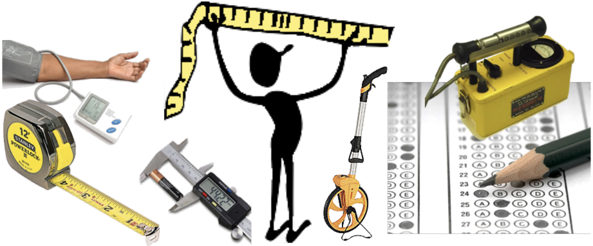
```
## [1] 159.3684
```

Histogram of tallest heights:



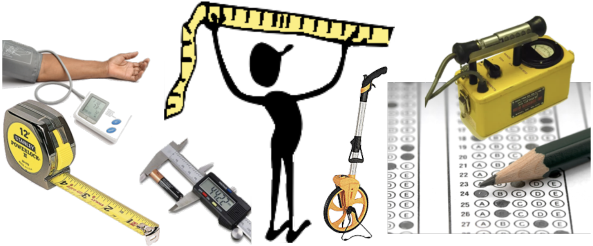
What's going on?

Source of error: Measurement



This is a common **source** of **error** which must **always be kept in mind**.

Source of error: Measurement

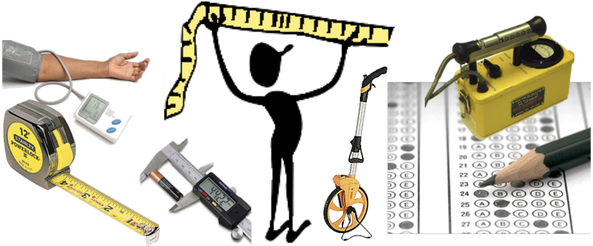


This is a common **source** of **error** which must **always be kept in mind**.

Examples:

- guessing the height of the tallest known redwood in metres

Source of error: Measurement



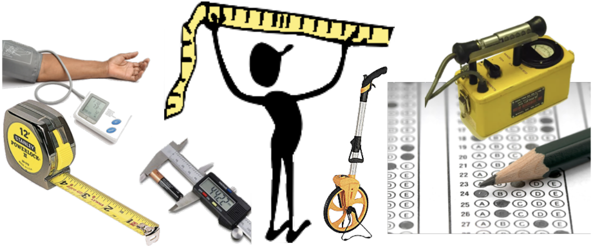
This is a common **source** of **error** which must **always be kept in mind**.

Examples:

- guessing the height of the tallest known redwood in metres
- even a binary measurement like informed consent from a patient to choose a treatment can have error

UNIVERSITY OF WATERLOO

Source of error: Measurement

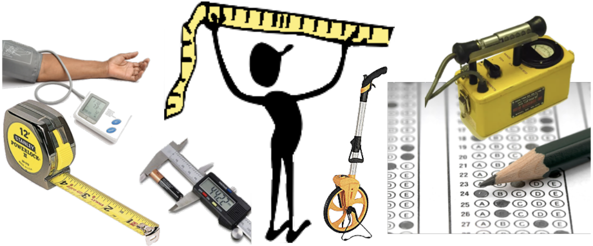


This is a common **source** of **error** which must **always be kept in mind**.

Examples:

- guessing the height of the tallest known redwood in metres
- even a binary measurement like informed consent from a patient to choose a treatment can have error
- the latitude and longitude of “Quebec” from Google

Source of error: Measurement

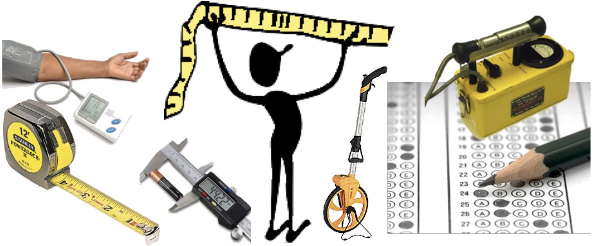


This is a common **source** of **error** which must **always be kept in mind**.

Examples:

- guessing the height of the tallest known redwood in metres
- even a binary measurement like informed consent from a patient to choose a treatment can have error
- the latitude and longitude of “Quebec” from Google
- think of which variates in `mtcars` might be most/least subject to measurement error

Source of error: Measurement



This is a common **source** of **error** which must **always be kept in mind**.

Examples:

- guessing the height of the tallest known redwood in metres
- even a binary measurement like informed consent from a patient to choose a treatment can have error
- the latitude and longitude of “Quebec” from Google
- think of which variates in `mtcars` might be most/least subject to measurement error
- the coordinates `x`, `y`, and `z` of `igg1` were

“ . . . determined by X-ray crystallography and as available to Padlan (1994) either from the Protein Data Bank or from original investigators at the time of publication.”

UNIVERSITY OF WATERLOO

Error, *bias, variability, and mean squared error*

By **error**, we mean a **single** instance, as in the difference between the measured value of a variate and its actual value (which may or may not be known).

***Error**, bias, variability, and mean squared error*

By **error**, we mean a **single** instance, as in the difference between the measured value of a variate and its actual value (which may or may not be known).

For example, in 2006 the tallest known redwood was discovered and found to be 115.7 metres (379.7 feet) tall. It was named "Hyperion", meaning "the high one", after one of the twelve Titan children of the Earth and the Sky from Greek mythology.

Each graduate student guess of this height might be thought of as a (very poor) measurement of Hyperion's height.

Error, *bias, variability, and mean squared error*

By **error**, we mean a **single** instance, as in the difference between the measured value of a variate and its actual value (which may or may not be known).

For example, in 2006 the tallest known redwood was discovered and found to be 115.7 metres (379.7 feet) tall. It was named "Hyperion", meaning "the high one", after one of the twelve Titan children of the Earth and the Sky from Greek mythology.

Each graduate student guess of this height might be thought of as a (very poor) measurement of Hyperion's height.

If the ith student's guess is x_i metres, then its **measurement error** in metres is

$$e_i = x_i - 115.7.$$

***Error**, bias, variability, and mean squared error*

By **error**, we mean a **single** instance, as in the difference between the measured value of a variate and its actual value (which may or may not be known).

For example, in 2006 the tallest known redwood was discovered and found to be 115.7 metres (379.7 feet) tall. It was named "Hyperion", meaning "the high one", after one of the twelve Titan children of the Earth and the Sky from Greek mythology.

Each graduate student guess of this height might be thought of as a (very poor) measurement of Hyperion's height.

If the ith student's guess is x_i metres, then its **measurement error** in metres is

$$e_i = x_i - 115.7.$$

And since all 38 students "measured" the same height, there are 38 measurements, and therefore 38 measurement errors.

***Error**, bias, variability, and mean squared error*

By **error**, we mean a **single** instance, as in the difference between the measured value of a variate and its actual value (which may or may not be known).

For example, in 2006 the tallest known redwood was discovered and found to be 115.7 metres (379.7 feet) tall. It was named "Hyperion", meaning "the high one", after one of the twelve Titan children of the Earth and the Sky from Greek mythology.

Each graduate student guess of this height might be thought of as a (very poor) measurement of Hyperion's height.

If the ith student's guess is x_i metres, then its **measurement error** in metres is

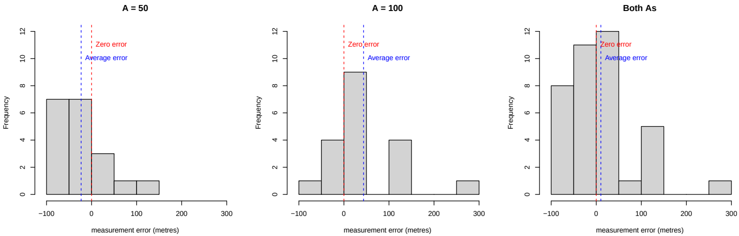
$$e_i = x_i - 115.7.$$

And since all 38 students "measured" the same height, there are 38 measurements, and therefore 38 measurement errors.

Note however that we have two measuring systems, one where the idea that the greatest height might be 50 metres was first planted, and one where the idea that the greatest height might be 100 metres was first planted.

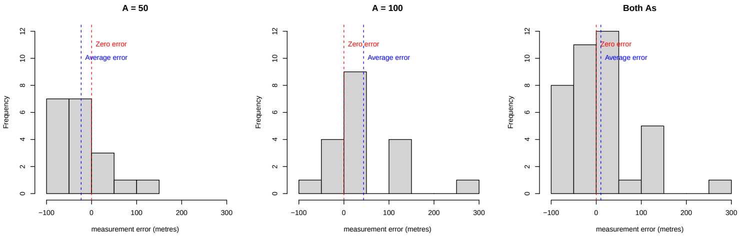
*Error, **bias**, variability, and mean squared error*

Histograms of these measurement errors for each group and then combined



*Error, **bias**, variability, and mean squared error*

Histograms of these measurement errors for each group and then combined



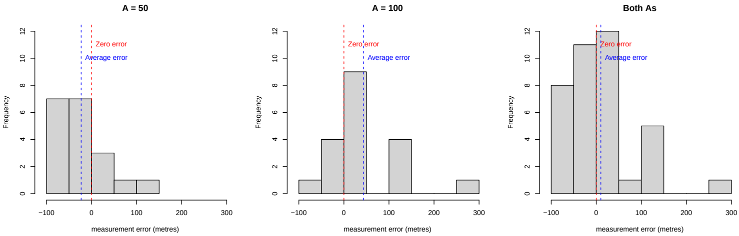
The average of all possible measurement errors is called the **measuring bias** given by

$$\frac{1}{N}\sum_{i\in\mathcal{P}} e_i = \frac{1}{N}\sum_{i\in\mathcal{P}}(x_i - x_{true}) = \overline{x} - x_{true}$$

where $\mathcal{P}$ is the population of size N containing all possible measurements of the same quantity (here $x_{true} = x_{Hyperion} = 115.7$ metres) from the same measuring system.

*Error, **bias**, variability, and mean squared error*

Histograms of these measurement errors for each group and then combined



The average of all possible measurement errors is called the **measuring bias** given by

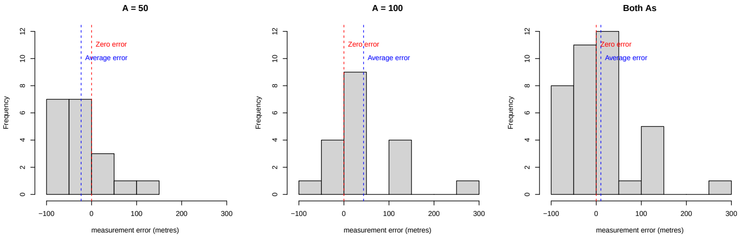
$$\frac{1}{N}\sum_{i\in\mathcal{P}} e_i = \frac{1}{N}\sum_{i\in\mathcal{P}}(x_i - x_{true}) = \overline{x} - x_{true}$$

where $\mathcal{P}$ is the population of size N containing all possible measurements of the same quantity (here $x_{true} = x_{Hyperion} = 115.7$ metres) from the same measuring system.

The histograms mark the **estimated measuring bias** based on the two separate samples, and the combined sample, containing 19, 19, and 38 measurement errors, respectively.

*Error, **bias**, variability, and mean squared error*

Histograms of these measurement errors for each group and then combined



The average of all possible measurement errors is called the **measuring bias** given by

$$\frac{1}{N}\sum_{i \in \mathcal{P}} e_i = \frac{1}{N}\sum_{i \in \mathcal{P}} (x_i - x_{true}) = \overline{x} - x_{true}$$

where $\mathcal{P}$ is the population of size N containing all possible measurements of the same quantity (here $x_{true} = x_{Hyperion} = 115.7$ metres) from the same measuring system.

The histograms mark the **estimated measuring bias** based on the two separate samples, and the combined sample, containing 19, 19, and 38 measurement errors, respectively.

Estimates of the measuring bias for each of the first, second, and then combined measuring systems are respectively, -23.2, 43.7, and 10.2 metres.

*Error, bias, **variability**, and mean squared error*

Similarly, the **measuring variability** of a measuring system can be defined as

$$\frac{1}{N}\sum_{i\in\mathcal{P}}(e_i-\overline{e})^2=\frac{1}{N}\sum_{i\in\mathcal{P}}(x_i-\overline{x})^2$$

with $\mathcal{P}$, N, and $\overline{x}$ defined as before.

Similarly, the **measuring variability** of a measuring system can be defined as

$$\frac{1}{N}\sum_{i\in\mathcal{P}}(e_i - \overline{e})^2 = \frac{1}{N}\sum_{i\in\mathcal{P}}(x_i - \overline{x})^2$$

with $\mathcal{P}$, N, and $\overline{x}$ defined as before.

For the three measuring systems, $A = 50$, $A = 100$, and the combined system, the estimates of the measuring variability are calculated to be (based on the available measurement errors for each) 3058.9, 6389.6, and 5743.7 squared metres, respectively.

*Error, bias, **variability**, and mean squared error*

Similarly, the **measuring variability** of a measuring system can be defined as

$$\frac{1}{N}\sum_{i\in\mathcal{P}}(e_i-\overline{e})^2=\frac{1}{N}\sum_{i\in\mathcal{P}}(x_i-\overline{x})^2$$

with $\mathcal{P}$, N, and $\overline{x}$ defined as before.

For the three measuring systems, $A = 50$, $A = 100$, and the combined system, the estimates of the measuring variability are calculated to be (based on the available measurement errors for each) 3058.9, 6389.6, and 5743.7 squared metres, respectively. (Note that these calculations replaced N by $n-1$ since the average error must be estimated.)

Similarly, the **measuring variability** of a measuring system can be defined as

$$\frac{1}{N}\sum_{i\in\mathcal{P}}(e_i-\overline{e})^2=\frac{1}{N}\sum_{i\in\mathcal{P}}(x_i-\overline{x})^2$$

with $\mathcal{P}$, N, and $\overline{x}$ defined as before.

For the three measuring systems, $A=50$, $A=100$, and the combined system, the estimates of the measuring variability are calculated to be (based on the available measurement errors for each) 3058.9, 6389.6, and 5743.7 squared metres, respectively. (Note that these calculations replaced N by $n-1$ since the average error must be estimated.)

Expressing these as estimated **standard deviations** of the measuring systems (i.e. by taking the square roots) gives quantities on the same scale as the errors, namely 55.3, 79.9, and 75.8 metres.

Error, bias, variability, and ***mean squared error***

Similarly, the **measuring mean squared error** of a measuring system can be defined as

$$\frac{1}{N}\sum_{i\in\mathcal{P}}(e_i - 0)^2 = \frac{1}{N}\sum_{i\in\mathcal{P}}(x_i - x_{true})^2$$

which can be shown to be the sum of the measuring variability and the square of the measuring bias. (Exercise: prove this.)

Error, bias, variability, and ***mean squared error***

Similarly, the **measuring mean squared error** of a measuring system can be defined as

$$\frac{1}{N}\sum_{i\in\mathcal{P}}(e_i - 0)^2 = \frac{1}{N}\sum_{i\in\mathcal{P}}(x_i - x_{true})^2$$

which can be shown to be the sum of the measuring variability and the square of the measuring bias. (Exercise: prove this.)

Thus a mean squared error always combines the variability and the bias (squared) into a single overall measure of the accuracy.

Error, bias, variability, and ***mean squared error***

Similarly, the **measuring mean squared error** of a measuring system can be defined as

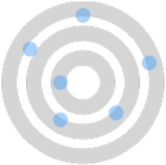
$$\frac{1}{N}\sum_{i\in\mathcal{P}}(e_i-0)^2=\frac{1}{N}\sum_{i\in\mathcal{P}}(x_i-x_{true})^2$$

which can be shown to be the sum of the measuring variability and the square of the measuring bias. (Exercise: prove this.)

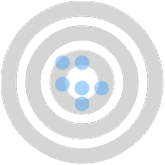
Thus a mean squared error always combines the variability and the bias (squared) into a single overall measure of the accuracy.

Their relation may be illustrated using targets:

low variability
high bias



high variability
low bias



low mean
squared error

Error, bias, variability, and ***mean squared error***

Similarly, the **measuring mean squared error** of a measuring system can be defined as

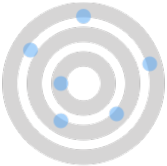
$$\frac{1}{N}\sum_{i \in \mathcal{P}} (e_i - 0)^2 = \frac{1}{N}\sum_{i \in \mathcal{P}} (x_i - x_{true})^2$$

which can be shown to be the sum of the measuring variability and the square of the measuring bias. (Exercise: prove this.)

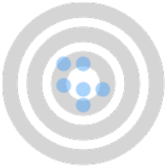
Thus a mean squared error always combines the variability and the bias (squared) into a single overall measure of the accuracy.

Their relation may be illustrated using targets:

low variability
high bias



high variability
low bias



low mean
squared error

There is often a trade off between variability and bias. Most practitioners prefer to improve (i.e. lower) the variability of a measuring system, then, afterwards, to reduce the bias.

Measuring system components

Measuring systems always have **at least three components** which can produce error:

(a) the **gauge**/instrument being used

Measuring system components

Measuring systems always have **at least three components** which can produce error:

(a) the **gauge**/instrument being used

(b) the **operator**/person doing and recording the measuring

Measuring system components

Measuring systems always have **at least three components** which can produce error:

(a) the **gauge**/instrument being used

(b) the **operator**/person doing and recording the measuring

(c) the **method** used to conduct the measuring (e.g. how is the object presented to be measured)

Sources of error

Recall that what is really of interest is a good reliable **population attribute** $a(\mathcal{P})$.

Sources of error

Recall that what is really of interest is a good reliable **population attribute** $a(\mathcal{P})$.

And population attributes are functions of the **units** of the population as well as the **variates**

Sources of error

Recall that what is really of interest is a good reliable **population attribute** $a(\mathcal{P})$.

And population attributes are functions of the **units** of the population as well as the **variates** Measuring values of variates is therefore only one possible source of error.

Sources of error

Recall that what is really of interest is a good reliable **population attribute** $a(\mathcal{P})$.

And population attributes are functions of the **units** of the population as well as the **variates** Measuring values of variates is therefore only one possible source of error. It might even be the least important.

Sources of error

Recall that what is really of interest is a good reliable **population attribute** $a(\mathcal{P})$.

And population attributes are functions of the **units** of the population as well as the **variates** Measuring values of variates is therefore only one possible source of error. It might even be the least important.

Focusing on the $\mathcal{P}$ and the units u which make it up, for each of the following think about what makes up the $\mathcal{P}$ we have in hand and contrast it with the $\mathcal{P}$ we are actually interested in:

- the `minority` data from the 2006 census on 33 census metropolitan areas,

UNIVERSITY OF WATERLOO

Sources of error

Recall that what is really of interest is a good reliable **population attribute** $a(\mathcal{P})$.

And population attributes are functions of the **units** of the population as well as the **variates** Measuring values of variates is therefore only one possible source of error. It might even be the least important.

Focusing on the $\mathcal{P}$ and the units u which make it up, for each of the following think about what makes up the $\mathcal{P}$ we have in hand and contrast it with the $\mathcal{P}$ we are actually interested in:

- the `minority` data from the 2006 census on 33 census metropolitan areas,
- the `mtcars` data on 32 cars (1973–74 models) extracted from issues of the U.S. magazine *Motor Trend* appearing in 1974,

Sources of error

Recall that what is really of interest is a good reliable **population attribute** $a(\mathcal{P})$.

And population attributes are functions of the **units** of the population as well as the **variates** Measuring values of variates is therefore only one possible source of error. It might even be the least important.

Focusing on the $\mathcal{P}$ and the units u which make it up, for each of the following think about what makes up the $\mathcal{P}$ we have in hand and contrast it with the $\mathcal{P}$ we are actually interested in:

- the `minority` data from the 2006 census on 33 census metropolitan areas,
- the `mtcars` data on 32 cars (1973–74 models) extracted from issues of the U.S. magazine *Motor Trend* appearing in 1974,
- the `igg1` data on 1556 alpha carbons in the human immunoglobulin G1 molecule,

UNIVERSITY OF WATERLOO

Sources of error

Recall that what is really of interest is a good reliable **population attribute** $a(\mathcal{P})$.

And population attributes are functions of the **units** of the population as well as the **variates** Measuring values of variates is therefore only one possible source of error. It might even be the least important.

Focusing on the $\mathcal{P}$ and the units u which make it up, for each of the following think about what makes up the $\mathcal{P}$ we have in hand and contrast it with the $\mathcal{P}$ we are actually interested in:

- the `minority` data from the 2006 census on 33 census metropolitan areas,
- the `mtcars` data on 32 cars (1973–74 models) extracted from issues of the U.S. magazine *Motor Trend* appearing in 1974,
- the `igg1` data on 1556 alpha carbons in the human immunoglobulin G1 molecule,
- the data on people's preferences for surgery or radiation,

Sources of error

Recall that what is really of interest is a good reliable **population attribute** $a(\mathcal{P})$.

And population attributes are functions of the **units** of the population as well as the **variates** Measuring values of variates is therefore only one possible source of error. It might even be the least important.

Focusing on the $\mathcal{P}$ and the units u which make it up, for each of the following think about what makes up the $\mathcal{P}$ we have in hand and contrast it with the $\mathcal{P}$ we are actually interested in:

- the `minority` data from the 2006 census on 33 census metropolitan areas,
- the `mtcars` data on 32 cars (1973–74 models) extracted from issues of the U.S. magazine *Motor Trend* appearing in 1974,
- the `igg1` data on 1556 alpha carbons in the human immunoglobulin G1 molecule,
- the data on people's preferences for surgery or radiation, or
- the data on people's guesses for the height of the tallest known redwood.

UNIVERSITY OF WATERLOO

Target populations, study populations, and samples

There are always at least three distinct conceptual sets of units in any study:

1. The **target population**, $\mathcal{P}_{Target}$.
   - the population of units u about whose population attribute(s) $a(\mathcal{P}_{Target})$ we truly want to know

Target populations, study populations, and samples

There are always at least three distinct conceptual sets of units in any study:

1. The **target population**, $\mathcal{P}_{Target}$.
   - the population of units u about whose population attribute(s) $a(\mathcal{P}_{Target})$ we truly want to know
2. The **study population**, $\mathcal{P}_{Study}$.
   - the population of units u which are possible to access/select

Target populations, study populations, and samples

There are always at least three distinct conceptual sets of units in any study:

1. The **target population**, $\mathcal{P}_{Target}$.
   - the population of units u about whose population attribute(s) $a(\mathcal{P}_{Target})$ we truly want to know
2. The **study population**, $\mathcal{P}_{Study}$.
   - the population of units u which are possible to access/select
   - any attribute $a(\mathcal{P}_{Study})$ is **surrogate** for the corresponding $a(\mathcal{P}_{Target})$

Target populations, study populations, and samples

There are always at least three distinct conceptual sets of units in any study:

1. The **target population**, $\mathcal{P}_{Target}$.
   - the population of units u about whose population attribute(s) $a(\mathcal{P}_{Target})$ we truly want to know
2. The **study population**, $\mathcal{P}_{Study}$.
   - the population of units u which are possible to access/select
   - any attribute $a(\mathcal{P}_{Study})$ is **surrogate** for the corresponding $a(\mathcal{P}_{Target})$
3. The **sample**, $\mathcal{S} \subset \mathcal{P}_{Study}$.
   - this is **not** a population

Target populations, study populations, and samples

There are always at least three distinct conceptual sets of units in any study:

1. The **target population**, $\mathcal{P}_{Target}$.
   - the population of units u about whose population attribute(s) $a(\mathcal{P}_{Target})$ we truly want to know
2. The **study population**, $\mathcal{P}_{Study}$.
   - the population of units u which are possible to access/select
   - any attribute $a(\mathcal{P}_{Study})$ is **surrogate** for the corresponding $a(\mathcal{P}_{Target})$
3. The **sample**, $\mathcal{S} \subset \mathcal{P}_{Study}$.
   - this is **not** a population but rather a set of units u which have **actually been selected** from those available in $\mathcal{P}_{Study}$

UNIVERSITY OF WATERLOO

Target populations, study populations, and samples

There are always at least three distinct conceptual sets of units in any study:

1. The **target population**, $\mathcal{P}_{Target}$.
   - the population of units u about whose population attribute(s) $a(\mathcal{P}_{Target})$ we truly want to know
2. The **study population**, $\mathcal{P}_{Study}$.
   - the population of units u which are possible to access/select
   - any attribute $a(\mathcal{P}_{Study})$ is **surrogate** for the corresponding $a(\mathcal{P}_{Target})$
3. The **sample**, $\mathcal{S} \subset \mathcal{P}_{Study}$.
   - this is **not** a population but rather a set of units u which have **actually been selected** from those available in $\mathcal{P}_{Study}$
   - the collection $u \in \mathcal{S}$ are the only u we actually have from $\mathcal{P}_{Study}$
   - any attribute $a(\mathcal{S})$ is **surrogate** for $a(\mathcal{P}_{Study})$

Target populations, study populations, and samples

For example, suppose the **target population** is all people in Canada now.

Target populations, study populations, and samples

For example, suppose the **target population** is all people in Canada now.Maybe 35 million people in $\mathcal{P}_{Target}$

Target populations, study populations, and samples

For example, suppose the **target population** is all people in Canada now. Maybe 35 million people in $\mathcal{P}_{Target}$

The **study population** is all people available for us to study now

Target populations, study populations, and samples

For example, suppose the **target population** is all people in Canada now. Maybe 35 million people in $\mathcal{P}_{Target}$

The **study population** is all people available for us to study now (e.g. all full time registered students presently here at the university).

Target populations, study populations, and samples

For example, suppose the **target population** is all people in Canada now.Maybe 35 million people in $\mathcal{P}_{Target}$

The **study population** is all people available for us to study now (e.g. all full time registered students presently here at the university). Maybe 35 thousand people in $\mathcal{P}_{Study}$.

Target populations, study populations, and samples

For example, suppose the **target population** is all people in Canada now.Maybe 35 million people in $\mathcal{P}_{Target}$

The **study population** is all people available for us to study now (e.g. all full time registered students presently here at the university). Maybe 35 thousand people in $\mathcal{P}_{Study}$.

The **sample** is the set of all students present in class today.

Target populations, study populations, and samples

For example, suppose the **target population** is all people in Canada now.Maybe 35 million people in $\mathcal{P}_{Target}$

The **study population** is all people available for us to study now (e.g. all full time registered students presently here at the university). Maybe 35 thousand people in $\mathcal{P}_{Study}$.

The **sample** is the set of all students present in class today. Maybe 35 people in $\mathcal{S}$.

Target populations, study populations, and samples

For example, suppose the **target population** is all people in Canada now.Maybe 35 million people in $\mathcal{P}_{Target}$

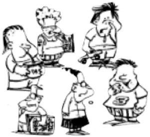
The **study population** is all people available for us to study now (e.g. all full time registered students presently here at the university). Maybe 35 thousand people in $\mathcal{P}_{Study}$.

The **sample** is the set of all students present in class today. Maybe 35 people in $\mathcal{S}$.

In pictures, we have something like

Target population
$\mathcal{P}_{Target}$

Study Population
$\mathcal{P}_{Study}$



Sample
$\mathcal{S}$

UNIVERSITY OF WATERLOO

Target populations, study populations, and samples

Or looking at how we might draw conclusions:

Target population
$\mathcal{P}_{Target}$

Study Population
$\mathcal{P}_{Study}$

Sample
$\mathcal{S}$

Target populations, study populations, and samples

Or worse, and fairly common in medical studies:

Target population
$\mathcal{P}_{Target}$

Study Population
$\mathcal{P}_{Study}$

Sample
$\mathcal{S}$

Sources of error

We are ultimately trying to infer attributes for the target population from the attribute values we find on the sample.

And there could be error in this inference!

Sources of error

We are ultimately trying to infer attributes for the target population from the attribute values we find on the sample.

And there could be error in this inference!

It is useful to separate that error as

$$\text{Inference Error} \quad =$$

Sources of error

We are ultimately trying to infer attributes for the target population from the attribute values we find on the sample.

And there could be error in this inference!

It is useful to separate that error as

$$\text{Inference Error} \quad = \quad a(\mathcal{S}) - a(\mathcal{P}_{Target})$$

UNIVERSITY OF WATERLOO

Sources of error

We are ultimately trying to infer attributes for the target population from the attribute values we find on the sample.

And there could be error in this inference!

It is useful to separate that error as

$$\begin{aligned} \text{Inference Error} &= a(\mathcal{S}) - a(\mathcal{P}_{Target}) \\ &= \{a(\mathcal{P}_{Study}) - a(\mathcal{P}_{Target})\} \end{aligned}$$

Sources of error

We are ultimately trying to infer attributes for the target population from the attribute values we find on the sample.

And there could be error in this inference!

It is useful to separate that error as

$$\begin{aligned}\text{Inference Error} &= a(\mathcal{S}) - a(\mathcal{P}_{Target}) \\ &= \{a(\mathcal{P}_{Study}) - a(\mathcal{P}_{Target})\} \\ &\qquad + \quad \{a(\mathcal{S}) - a(\mathcal{P}_{Study})\}\end{aligned}$$

Sources of error

We are ultimately trying to infer attributes for the target population from the attribute values we find on the sample.

And there could be error in this inference!

It is useful to separate that error as

$$\begin{aligned}
\text{Inference Error} &= a(\mathcal{S}) - a(\mathcal{P}_{Target}) \\
&= \{a(\mathcal{P}_{Study}) - a(\mathcal{P}_{Target})\} \\
&\qquad + \{a(\mathcal{S}) - a(\mathcal{P}_{Study})\}
\end{aligned}$$

The first term relates the attribute's value on the **study** population to its value on the **target** population;

Sources of error

We are ultimately trying to infer attributes for the target population from the attribute values we find on the sample.

And there could be error in this inference!

It is useful to separate that error as

$$\begin{aligned}\text{Inference Error} &= a(\mathcal{S}) - a(\mathcal{P}_{Target}) \\ &= \{a(\mathcal{P}_{Study}) - a(\mathcal{P}_{Target})\} \\ &\quad + \{a(\mathcal{S}) - a(\mathcal{P}_{Study})\}\end{aligned}$$

The first term relates the attribute's value on the **study** population to its value on the **target** population; the second the attribute's value on the **sample** to its value on the **study** population.

Sources of error

We are ultimately trying to infer attributes for the target population from the attribute values we find on the sample.

And there could be error in this inference!

It is useful to separate that error as

$$\begin{aligned}\text{Inference Error} &= a(\mathcal{S}) - a(\mathcal{P}_{Target}) \\ &= \{a(\mathcal{P}_{Study}) - a(\mathcal{P}_{Target})\} \\ &\quad + \{a(\mathcal{S}) - a(\mathcal{P}_{Study})\}\end{aligned}$$

The first term relates the attribute's value on the **study** population to its value on the **target** population; the second the attribute's value on the **sample** to its value on the **study** population.

Separation like this allows us to focus on where the error sources might be and what might be done about them.

UNIVERSITY OF WATERLOO

Study error

The **study error** is defined to be

$$\text{Study error} = a(\mathcal{P}_{Study}) - a(\mathcal{P}_{Target})$$

Study error

The **study error** is defined to be

$$\text{Study error} = a(\mathcal{P}_{Study}) - a(\mathcal{P}_{Target})$$

This error necessarily depends on

- the attribute $a(\cdot)$,

Study error

The **study error** is defined to be

$$\text{Study error} = a(\mathcal{P}_{Study}) - a(\mathcal{P}_{Target})$$

This error necessarily depends on

- the attribute $a(\cdot)$,
- the study population $\mathcal{P}_{Study}$,

Study error

The **study error** is defined to be

$$\text{Study error} = a(\mathcal{P}_{Study}) - a(\mathcal{P}_{Target})$$

This error necessarily depends on

- the attribute $a(\cdot)$,
- the study population $\mathcal{P}_{Study}$, and
- the target population $\mathcal{P}_{Target}$.

Study error

The **study error** is defined to be

$$\text{Study error} = a(\mathcal{P}_{Study}) - a(\mathcal{P}_{Target})$$

This error necessarily depends on

- the attribute $a(\cdot)$,
- the study population $\mathcal{P}_{Study}$, and
- the target population $\mathcal{P}_{Target}$.

Making the case that this error is small, or ignorable, can be difficult and often must be made on extra-statistical grounds.

Study error

The **study error** is defined to be

$$\text{Study error} = a(\mathcal{P}_{Study}) - a(\mathcal{P}_{Target})$$

This error necessarily depends on

- the attribute $a(\cdot)$,
- the study population $\mathcal{P}_{Study}$, and
- the target population $\mathcal{P}_{Target}$.

Making the case that this error is small, or ignorable, can be difficult and often must be made on extra-statistical grounds.

And the size of the error may never be known, at least at the time of the study (e.g. $\mathcal{P}_{Target}$ contains *units* from the future).

Study error

The **study error** is defined to be

$$\text{Study error} = a(\mathcal{P}_{Study}) - a(\mathcal{P}_{Target})$$

This error necessarily depends on

- the attribute $a(\cdot)$,
- the study population $\mathcal{P}_{Study}$, and
- the target population $\mathcal{P}_{Target}$.

Making the case that this error is small, or ignorable, can be difficult and often must be made on extra-statistical grounds.

And the size of the error may never be known, at least at the time of the study (e.g. $\mathcal{P}_{Target}$ contains *units* from the future).

Note that

- the study error could be small even if $\mathcal{P}_{Study}$ is very different from $\mathcal{P}_{Target}$

Study error

The **study error** is defined to be

$$\text{Study error} = a(\mathcal{P}_{Study}) - a(\mathcal{P}_{Target})$$

This error necessarily depends on

- the attribute $a(\cdot)$,
- the study population $\mathcal{P}_{Study}$, and
- the target population $\mathcal{P}_{Target}$.

Making the case that this error is small, or ignorable, can be difficult and often must be made on extra-statistical grounds.

And the size of the error may never be known, at least at the time of the study (e.g. $\mathcal{P}_{Target}$ contains *units* from the future).

Note that

- the study error could be small even if $\mathcal{P}_{Study}$ is very different from $\mathcal{P}_{Target}$ depending on what the attribute $a()$ is,

UNIVERSITY OF WATERLOO

Study error

The **study error** is defined to be

$$\text{Study error} = a(\mathcal{P}_{Study}) - a(\mathcal{P}_{Target})$$

This error necessarily depends on

- the attribute $a(\cdot)$,
- the study population $\mathcal{P}_{Study}$, and
- the target population $\mathcal{P}_{Target}$.

Making the case that this error is small, or ignorable, can be difficult and often must be made on extra-statistical grounds.

And the size of the error may never be known, at least at the time of the study (e.g. $\mathcal{P}_{Target}$ contains *units* from the future).

Note that

- the study error could be small even if $\mathcal{P}_{Study}$ is very different from $\mathcal{P}_{Target}$ depending on what the attribute $a()$ is, or,
- the study error could be large even when we could argue that $\mathcal{P}_{Target}$ should be very much like $\mathcal{P}_{Study}$, again depending on the attribute (e.g. $a(\mathcal{P}) = max_{u \in \mathcal{P}} X(u)$ and $\mathcal{P}_{Study} \subset \mathcal{P}_{Target}$).

Sample error

Similarly, the **sample error** is defined to be

$$\text{Sample error} = a(\mathcal{S}) - a(\mathcal{P}_{Study})$$

Sample error

Similarly, the **sample error** is defined to be

$$\text{Sample error} = a(\mathcal{S}) - a(\mathcal{P}_{Study})$$

This error also necessarily depends on

- the attribute $a(\cdot)$,

Sample error

Similarly, the **sample error** is defined to be

$$\text{Sample error} = a(\mathcal{S}) - a(\mathcal{P}_{Study})$$

This error also necessarily depends on

- the attribute $a(\cdot)$,
- the study population $\mathcal{P}_{Study}$,

Sample error

Similarly, the **sample error** is defined to be

$$\text{Sample error} = a(\mathcal{S}) - a(\mathcal{P}_{Study})$$

This error also necessarily depends on

- the attribute $a(\cdot)$,
- the study population $\mathcal{P}_{Study}$, and
- the sample $\mathcal{S}$.

UNIVERSITY OF WATERLOO

Sample error

Similarly, the **sample error** is defined to be

$$\text{Sample error} = a(\mathcal{S}) - a(\mathcal{P}_{Study})$$

This error also necessarily depends on

- the attribute $a(\cdot)$,
- the study population $\mathcal{P}_{Study}$, and
- the sample $\mathcal{S}$.

A fundamental and important difference between sample error and study error is that

- $\mathcal{S} \subset \mathcal{P}_{Study}$,

Sample error

Similarly, the **sample error** is defined to be

$$\text{Sample error} = a(\mathcal{S}) - a(\mathcal{P}_{Study})$$

This error also necessarily depends on

- the attribute $a(\cdot)$,
- the study population $\mathcal{P}_{Study}$, and
- the sample $\mathcal{S}$.

A fundamental and important difference between sample error and study error is that

- $\mathcal{S} \subset \mathcal{P}_{Study}$,
- and it could be that both
  - the $\mathcal{P}_{Study} \subset \mathcal{P}_{Target}$

Sample error

Similarly, the **sample error** is defined to be

$$\text{Sample error} = a(\mathcal{S}) - a(\mathcal{P}_{Study})$$

This error also necessarily depends on

- the attribute $a(\cdot)$,
- the study population $\mathcal{P}_{Study}$, and
- the sample $\mathcal{S}$.

A fundamental and important difference between sample error and study error is that

- $\mathcal{S} \subset \mathcal{P}_{Study}$,
- and it could be that both
  - the $\mathcal{P}_{Study} \subset \mathcal{P}_{Target}$ and
  - $\mathcal{P}_{Study} \not\subset \mathcal{P}_{Target}$

  are possible.

Sample error

Similarly, the **sample error** is defined to be

$$\text{Sample error} = a(\mathcal{S}) - a(\mathcal{P}_{Study})$$

This error also necessarily depends on

- the attribute $a(\cdot)$,
- the study population $\mathcal{P}_{Study}$, and
- the sample $\mathcal{S}$.

A fundamental and important difference between sample error and study error is that

- $\mathcal{S} \subset \mathcal{P}_{Study}$,
- and it could be that both
  - the $\mathcal{P}_{Study} \subset \mathcal{P}_{Target}$ and
  - $\mathcal{P}_{Study} \not\subset \mathcal{P}_{Target}$

  are possible.

E.g. When $\mathcal{P}_{Target}$ consists of humans, having $\mathcal{P}_{Study} \subset \mathcal{P}_{Target}$ could be unethical.

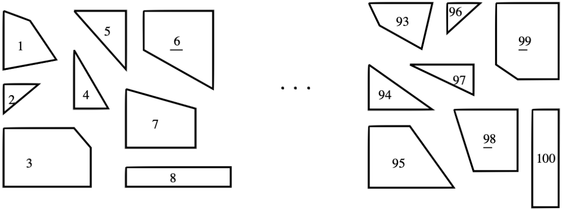
Sample selection – the blocks competition

Consider a study population $\mathcal{P}_{Study}$ consisting of $N = 100$ blocks labelled $u = 1, 2, 3, \ldots, 100$.

Sample selection – the blocks competition

Consider a study population $\mathcal{P}_{Study}$ consisting of $N = 100$ blocks labelled $u = 1, 2, 3, \ldots, 100$.

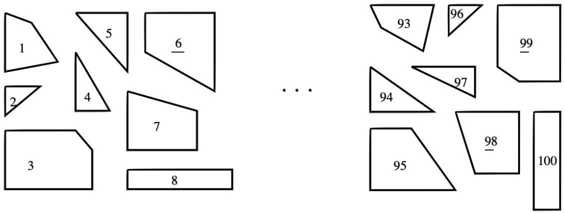
The blocks are of of uniform thickness and density (all blocks were cut from the same opaque plastic sheet of about 5mm thickness), but have different shapes as such as shown below:



Sample selection – the blocks competition

Consider a study population $\mathcal{P}_{Study}$ consisting of $N = 100$ blocks labelled $u = 1, 2, 3, \ldots, 100$.

The blocks are of of uniform thickness and density (all blocks were cut from the same opaque plastic sheet of about 5mm thickness), but have different shapes as such as shown below:

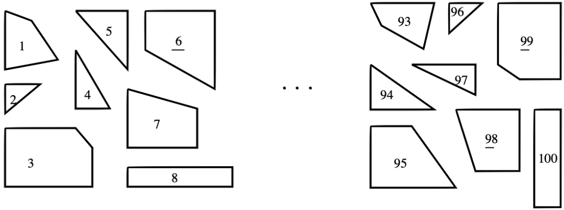


Suppose also that $\mathcal{P}_{Target} = \mathcal{P}_{Study}$ and that the attribute of interest is the average weight of all $N = 100$ blocks in the population.

Sample selection – the blocks competition

Consider a study population $\mathcal{P}_{Study}$ consisting of $N = 100$ blocks labelled $u = 1, 2, 3, \ldots, 100$.

The blocks are of of uniform thickness and density (all blocks were cut from the same opaque plastic sheet of about 5mm thickness), but have different shapes as such as shown below:



Suppose also that $\mathcal{P}_{Target} = \mathcal{P}_{Study}$ and that the attribute of interest is the average weight of all $N = 100$ blocks in the population.

We want a sample $\mathcal{S} \subset \mathcal{P}_{Study}$ of $n = 10$ blocks selected from the 100, whose average weight is (nearly) the same as the average weight of all 100.

That is, we would like a sample with zero (or at least small in absolute value) **sample error** $a(\mathcal{S}) - a(\mathcal{P}_{Study})$.

Sample selection – the blocks competition

The competition:

Sample selection – the blocks competition

The competition:

- Each person is to select a sample $\mathcal{S}$ of exactly **10 blocks** from the population $\mathcal{P}_{Study}$ of 100 blocks.

Sample selection – the blocks competition

The competition:

- Each person is to select a sample $\mathcal{S}$ of exactly **10 blocks** from the population $\mathcal{P}_{Study}$ of 100 blocks.
- The sample $\mathcal{S}$ is to be selected so that the average weight of the sample blocks **matches as closely as possible** to the average weight of all 100 blocks.

UNIVERSITY OF WATERLOO

Sample selection – the blocks competition

The competition:

- Each person is to select a sample $\mathcal{S}$ of exactly **10 blocks** from the population $\mathcal{P}_{Study}$ of 100 blocks.
- The sample $\mathcal{S}$ is to be selected so that the average weight of the sample blocks **matches as closely as possible** to the average weight of all 100 blocks.
- Person whose sample average weight is closest to the average for all 100 wins!

Sample selection – the blocks competition

The competition:

- Each person is to select a sample $\mathcal{S}$ of exactly **10 blocks** from the population $\mathcal{P}_{Study}$ of 100 blocks.
- The sample $\mathcal{S}$ is to be selected so that the average weight of the sample blocks **matches as closely as possible** to the average weight of all 100 blocks.
- Person whose sample average weight is closest to the average for all 100 wins!

The data collection process:

Sample selection – the blocks competition

The competition:

- Each person is to select a sample $\mathcal{S}$ of exactly **10 blocks** from the population $\mathcal{P}_{Study}$ of 100 blocks.
- The sample $\mathcal{S}$ is to be selected so that the average weight of the sample blocks **matches as closely as possible** to the average weight of all 100 blocks.
- Person whose sample average weight is closest to the average for all 100 wins!

The data collection process:

- Write your student id number on the file card.

Sample selection – the blocks competition

The competition:

- Each person is to select a sample $\mathcal{S}$ of exactly **10 blocks** from the population $\mathcal{P}_{Study}$ of 100 blocks.
- The sample $\mathcal{S}$ is to be selected so that the average weight of the sample blocks **matches as closely as possible** to the average weight of all 100 blocks.
- Person whose sample average weight is closest to the average for all 100 wins!

The data collection process:

- Write your student id number on the file card.
- Go examine the set of 100 blocks. **Do not touch them.** Just look.

Sample selection – the blocks competition

The competition:

- Each person is to select a sample $\mathcal{S}$ of exactly **10 blocks** from the population $\mathcal{P}_{Study}$ of 100 blocks.
- The sample $\mathcal{S}$ is to be selected so that the average weight of the sample blocks **matches as closely as possible** to the average weight of all 100 blocks.
- Person whose sample average weight is closest to the average for all 100 wins!

The data collection process:

- Write your student id number on the file card.
- Go examine the set of 100 blocks. **Do not touch them.** Just look.
- Choose 10 different blocks you think have an average weight close to that of all 100. Again, **no touching** the blocks.

Sample selection – the blocks competition

The competition:

- Each person is to select a sample $\mathcal{S}$ of exactly **10 blocks** from the population $\mathcal{P}_{Study}$ of 100 blocks.
- The sample $\mathcal{S}$ is to be selected so that the average weight of the sample blocks **matches as closely as possible** to the average weight of all 100 blocks.
- Person whose sample average weight is closest to the average for all 100 wins!

The data collection process:

- Write your student id number on the file card.
- Go examine the set of 100 blocks. **Do not touch them.** Just look.
- Choose 10 different blocks you think have an average weight close to that of all 100. Again, **no touching** the blocks.
- On the card, record the block numbers (u) of the 10 blocks you have chosen.

Sample selection – the blocks competition

The competition:

- Each person is to select a sample $\mathcal{S}$ of exactly **10 blocks** from the population $\mathcal{P}_{Study}$ of 100 blocks.
- The sample $\mathcal{S}$ is to be selected so that the average weight of the sample blocks **matches as closely as possible** to the average weight of all 100 blocks.
- Person whose sample average weight is closest to the average for all 100 wins!

The data collection process:

- Write your student id number on the file card.
- Go examine the set of 100 blocks. **Do not touch them.** Just look.
- Choose 10 different blocks you think have an average weight close to that of all 100. Again, **no touching** the blocks.
- On the card, record the block numbers (u) of the 10 blocks you have chosen.
- Hand in your card when your sample selection is complete.

UNIVERSITY OF WATERLOO

Repeated sampling

Suppose we are considering only samples $\mathcal{S}$ of n distinct units $u \in \mathcal{P}_{Study}$. So each sample is of size n, typically with $n << N$, the size of the study population $\mathcal{P}_{Study}$.

Repeated sampling

Suppose we are considering only samples $\mathcal{S}$ of n distinct units $u \in \mathcal{P}_{Study}$. So each sample is of size n, typically with $n << N$, the size of the study population $\mathcal{P}_{Study}$.

There are exactly $M = \binom{N}{n}$ such possibly distinct subsets $\mathcal{S}$ of $\mathcal{P}_{Study}$.

Repeated sampling

Suppose we are considering only samples $\mathcal{S}$ of n distinct units $u \in \mathcal{P}_{Study}$. So each sample is of size n, typically with $n << N$, the size of the study population $\mathcal{P}_{Study}$.

There are exactly $M = \binom{N}{n}$ such possibly distinct subsets $\mathcal{S}$ of $\mathcal{P}_{Study}$.

Each sample $\mathcal{S}_i$, $i = 1, \ldots, M$ will have its own sample error

$$e_i = a(\mathcal{S}_i) - a(\mathcal{P}_{Study})$$

for any attribute $a()$, where i now indexes the possible samples.

Repeated sampling

Suppose we are considering only samples $\mathcal{S}$ of n distinct units $u \in \mathcal{P}_{Study}$. So each sample is of size n, typically with $n << N$, the size of the study population $\mathcal{P}_{Study}$.

There are exactly $M = \binom{N}{n}$ such possibly distinct subsets $\mathcal{S}$ of $\mathcal{P}_{Study}$.

Each sample $\mathcal{S}_i$, $i = 1, \ldots, M$ will have its own sample error

$$e_i = a(\mathcal{S}_i) - a(\mathcal{P}_{Study})$$

for any attribute $a()$, where i now indexes the possible samples.

For any collection $\mathcal{C}$ of $N_{\mathcal{C}}$ samples, we can calculate the average error and the variability of those samples for any numerical attribute.

Repeated sampling

Suppose we are considering only samples $\mathcal{S}$ of n distinct units $u \in \mathcal{P}_{Study}$. So each sample is of size n, typically with $n << N$, the size of the study population $\mathcal{P}_{Study}$.

There are exactly $M = \binom{N}{n}$ such possibly distinct subsets $\mathcal{S}$ of $\mathcal{P}_{Study}$.

Each sample $\mathcal{S}_i$, $i = 1, \ldots, M$ will have its own sample error

$$e_i = a(\mathcal{S}_i) - a(\mathcal{P}_{Study})$$

for any attribute $a()$, where i now indexes the possible samples.

For any collection $\mathcal{C}$ of $N_{\mathcal{C}}$ samples, we can calculate the average error and the variability of those samples for any numerical attribute. These are called the **sampling bias** and **sampling variability** of $\mathcal{C}$ for that attribute.

Repeated sampling

Suppose we are considering only samples $\mathcal{S}$ of n distinct units $u \in \mathcal{P}_{Study}$. So each sample is of size n, typically with $n << N$, the size of the study population $\mathcal{P}_{Study}$.

There are exactly $M = \binom{N}{n}$ such possibly distinct subsets $\mathcal{S}$ of $\mathcal{P}_{Study}$.

Each sample $\mathcal{S}_i$, $i = 1, \ldots, M$ will have its own sample error

$$e_i = a(\mathcal{S}_i) - a(\mathcal{P}_{Study})$$

for any attribute $a()$, where i now indexes the possible samples.

For any collection $\mathcal{C}$ of $N_{\mathcal{C}}$ samples, we can calculate the average error and the variability of those samples for any numerical attribute. These are called the **sampling bias** and **sampling variability** of $\mathcal{C}$ for that attribute. These can be written as

$$\text{Sampling bias} = \frac{1}{N_{\mathcal{C}}} \sum_{\mathcal{S} \in \mathcal{C}} a(\mathcal{S}) - a(\mathcal{P}_{Study})$$

Repeated sampling

Suppose we are considering only samples $\mathcal{S}$ of n distinct units $u \in \mathcal{P}_{Study}$. So each sample is of size n, typically with $n << N$, the size of the study population $\mathcal{P}_{Study}$.

There are exactly $M = \binom{N}{n}$ such possibly distinct subsets $\mathcal{S}$ of $\mathcal{P}_{Study}$.

Each sample $\mathcal{S}_i$, $i = 1, \ldots, M$ will have its own sample error

$$e_i = a(\mathcal{S}_i) - a(\mathcal{P}_{Study})$$

for any attribute $a()$, where i now indexes the possible samples.

For any collection $\mathcal{C}$ of $N_{\mathcal{C}}$ samples, we can calculate the average error and the variability of those samples for any numerical attribute. These are called the **sampling bias** and **sampling variability** of $\mathcal{C}$ for that attribute. These can be written as

$$\text{Sampling bias} = \frac{1}{N_{\mathcal{C}}} \sum_{\mathcal{S} \in \mathcal{C}} a(\mathcal{S}) - a(\mathcal{P}_{Study}) = \overline{a}_{\mathcal{C}} - a(\mathcal{P}_{Study})$$

Repeated sampling

Suppose we are considering only samples $\mathcal{S}$ of n distinct units $u \in \mathcal{P}_{Study}$. So each sample is of size n, typically with $n << N$, the size of the study population $\mathcal{P}_{Study}$.

There are exactly $M = \binom{N}{n}$ such possibly distinct subsets $\mathcal{S}$ of $\mathcal{P}_{Study}$.

Each sample $\mathcal{S}_i$, $i = 1, \ldots, M$ will have its own sample error

$$e_i = a(\mathcal{S}_i) - a(\mathcal{P}_{Study})$$

for any attribute $a()$, where i now indexes the possible samples.

For any collection $\mathcal{C}$ of $N_{\mathcal{C}}$ samples, we can calculate the average error and the variability of those samples for any numerical attribute. These are called the **sampling bias** and **sampling variability** of $\mathcal{C}$ for that attribute. These can be written as

$$\text{Sampling bias} = \frac{1}{N_{\mathcal{C}}} \sum_{\mathcal{S} \in \mathcal{C}} a(\mathcal{S}) - a(\mathcal{P}_{Study}) = \overline{a}_{\mathcal{C}} - a(\mathcal{P}_{Study})$$

and

$$\text{Sampling variability} = \frac{1}{N_{\mathcal{C}}} \sum_{\mathcal{S} \in \mathcal{C}} \left[a(\mathcal{S}) - \overline{a}_{\mathcal{C}} \right]^2 .$$

Repeated sampling

Take $\mathcal{C}$ to be the collection of samples of size n from $\mathcal{P}_{Study}$; without any loss of generality, suppose each sample $\mathcal{S}_i$ above appears $k_i \geq 0$ times in $\mathcal{C}$.

UNIVERSITY OF WATERLOO

Repeated sampling

Take $\mathcal{C}$ to be the collection of samples of size n from $\mathcal{P}_{Study}$; without any loss of generality, suppose each sample $\mathcal{S}_i$ above appears $k_i \geq 0$ times in $\mathcal{C}$. So, in this case, $N_{\mathcal{C}} = \sum_{i=1}^{M} k_i$.

Repeated sampling

Take $\mathcal{C}$ to be the collection of samples of size n from $\mathcal{P}_{Study}$; without any loss of generality, suppose each sample $\mathcal{S}_i$ above appears $k_i \geq 0$ times in $\mathcal{C}$. So, in this case, $N_{\mathcal{C}} = \sum_{i=1}^{M} k_i$.

The sampling bias and variability are now written as

Repeated sampling

Take $\mathcal{C}$ to be the collection of samples of size n from $\mathcal{P}_{Study}$; without any loss of generality, suppose each sample $\mathcal{S}_i$ above appears $k_i \geq 0$ times in $\mathcal{C}$. So, in this case, $N_{\mathcal{C}} = \sum_{i=1}^{M} k_i$.

The sampling bias and variability are now written as

$$\text{Sampling bias} = \frac{1}{N_{\mathcal{C}}} \sum_{i=1}^{M} k_i a(\mathcal{S}_i) - a(\mathcal{P}_{Study})$$

Repeated sampling

Take $\mathcal{C}$ to be the collection of samples of size n from $\mathcal{P}_{Study}$; without any loss of generality, suppose each sample $\mathcal{S}_i$ above appears $k_i \geq 0$ times in $\mathcal{C}$. So, in this case, $N_{\mathcal{C}} = \sum_{i=1}^{M} k_i$.

The sampling bias and variability are now written as

$$\text{Sampling bias} = \frac{1}{N_{\mathcal{C}}} \sum_{i=1}^{M} k_i a(\mathcal{S}_i) - a(\mathcal{P}_{Study})$$

and

$$\text{Sampling variability} = \frac{1}{N_{\mathcal{C}}} \sum_{i=1}^{M} k_i \left[a(\mathcal{S}_i) - \overline{a}_{\mathcal{C}}\right]^2 .$$

Repeated sampling

Take $\mathcal{C}$ to be the collection of samples of size n from $\mathcal{P}_{Study}$; without any loss of generality, suppose each sample $\mathcal{S}_i$ above appears $k_i \geq 0$ times in $\mathcal{C}$. So, in this case, $N_{\mathcal{C}} = \sum_{i=1}^{M} k_i$.

The sampling bias and variability are now written as

$$\text{Sampling bias} = \frac{1}{N_{\mathcal{C}}} \sum_{i=1}^{M} k_i a(\mathcal{S}_i) - a(\mathcal{P}_{Study})$$

and

$$\text{Sampling variability} = \frac{1}{N_{\mathcal{C}}} \sum_{i=1}^{M} k_i \left[a(\mathcal{S}_i) - \overline{a}_{\mathcal{C}} \right]^2 .$$

If we were to select m different samples from $\mathcal{C}$ with

$$Pr(\mathcal{S} = \mathcal{S}_i) = \frac{k_i}{N_{\mathcal{C}}}$$

then both sampling bias and variability could be estimated from the m values of $a(\mathcal{S})$. (Replace $N_{\mathcal{C}}$ by $m-1$ in the variability estimate.)

Repeated sampling

Every sampling plan will produce a collection $\mathcal{C}$ of possible samples

Repeated sampling

Every sampling plan will produce a collection $\mathcal{C}$ of possible samples and so plans can be compared by comparing their sampling bias and variability.

Repeated sampling

Every sampling plan will produce a collection $\mathcal{C}$ of possible samples and so plans can be compared by comparing their sampling bias and variability.

Many statistical sampling plans have been developed to produce collections $\mathcal{C}$ that have small (even zero) sampling bias and low variability for particular attribute(s) of interest.

Repeated sampling

Every sampling plan will produce a collection $\mathcal{C}$ of possible samples and so plans can be compared by comparing their sampling bias and variability.

Many statistical sampling plans have been developed to produce collections $\mathcal{C}$ that have small (even zero) sampling bias and low variability for particular attribute(s) of interest.

Whatever the plan, in practice we have only a single sample $\mathcal{S}$ whose sample error may be large or small.

Repeated sampling

Every sampling plan will produce a collection $\mathcal{C}$ of possible samples and so plans can be compared by comparing their sampling bias and variability.

Many statistical sampling plans have been developed to produce collections $\mathcal{C}$ that have small (even zero) sampling bias and low variability for particular attribute(s) of interest.

Whatever the plan, in practice we have only a single sample $\mathcal{S}$ whose sample error may be large or small. We have no way of knowing for sure.

Repeated sampling

Every sampling plan will produce a collection $\mathcal{C}$ of possible samples and so plans can be compared by comparing their sampling bias and variability.

Many statistical sampling plans have been developed to produce collections $\mathcal{C}$ that have small (even zero) sampling bias and low variability for particular attribute(s) of interest.

Whatever the plan, in practice we have only a single sample $\mathcal{S}$ whose sample error may be large or small. We have no way of knowing for sure.

However, by choosing samples at random from $\mathcal{C}$ we have some idea of the *operating characteristics* from the plan.

Repeated sampling

Every sampling plan will produce a collection $\mathcal{C}$ of possible samples and so plans can be compared by comparing their sampling bias and variability.

Many statistical sampling plans have been developed to produce collections $\mathcal{C}$ that have small (even zero) sampling bias and low variability for particular attribute(s) of interest.

Whatever the plan, in practice we have only a single sample $\mathcal{S}$ whose sample error may be large or small. We have no way of knowing for sure.

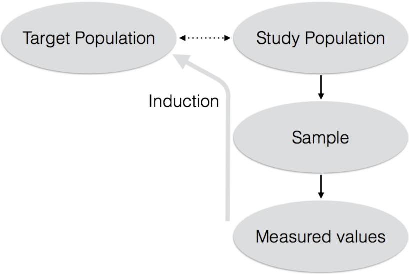
However, by choosing samples at random from $\mathcal{C}$ we have some idea of the *operating characteristics* from the plan. A plan with small (ideally zero) sampling bias and small sampling variability will have a small probability of producing a sample having a large sample error for that attribute.

Inductive inference

We draw conclusions about $a(\mathcal{P}_{Target})$ from an attribute $a_{\approx}(\mathcal{S}) = \widehat{a}(\mathcal{S})$, being based on measurements of units in some sample $\mathcal{S}$. A case needs to be made for the validity of each step along **the inductive path**:

Inductive inference

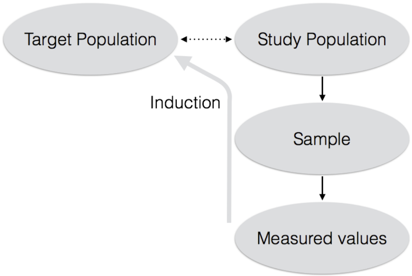
We draw conclusions about $a(\mathcal{P}_{Target})$ from an attribute $a_{\approx}(\mathcal{S}) = \widehat{a}(\mathcal{S})$, being based on measurements of units in some sample $\mathcal{S}$. A case needs to be made for the validity of each step along **the inductive path**:



UNIVERSITY OF WATERLOO

Inductive inference

We draw conclusions about $a(\mathcal{P}_{Target})$ from an attribute $a_{\approx}(\mathcal{S}) = \widehat{a}(\mathcal{S})$, being based on measurements of units in some sample $\mathcal{S}$. A case needs to be made for the validity of each step along **the inductive path**:

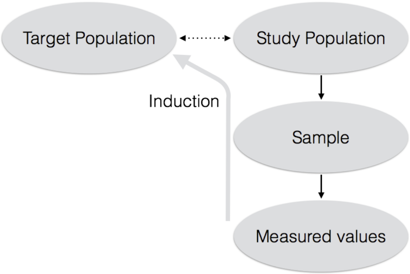


The overall **inductive error** separates into three conceptually distinct sources:

$$a_{\approx}(\mathcal{S}) - a(\mathcal{P}_{Target}) \quad =$$

Inductive inference

We draw conclusions about $a(\mathcal{P}_{Target})$ from an attribute $a_{\approx}(\mathcal{S}) = \widehat{a}(\mathcal{S})$, being based on measurements of units in some sample $\mathcal{S}$. A case needs to be made for the validity of each step along **the inductive path**:



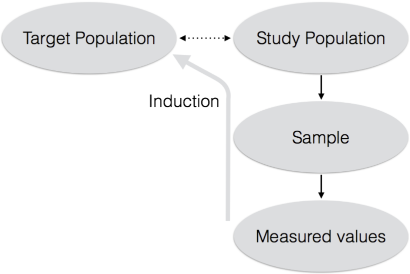
The overall **inductive error** separates into three conceptually distinct sources:

$$a_{\approx}(\mathcal{S}) - a(\mathcal{P}_{Target}) \quad = \quad [a_{\approx}(\mathcal{S}) - a(\mathcal{S})\ ] \qquad \ldots \quad \text{measurement}$$

UNIVERSITY OF WATERLOO

Inductive inference

We draw conclusions about $a(\mathcal{P}_{Target})$ from an attribute $a_{\approx}(\mathcal{S}) = \widehat{a}(\mathcal{S})$, being based on measurements of units in some sample $\mathcal{S}$. A case needs to be made for the validity of each step along **the inductive path**:



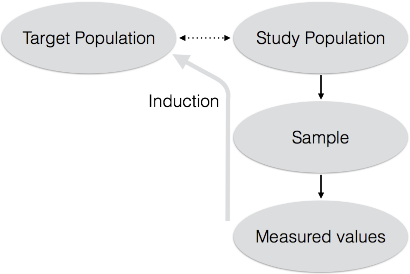
The overall **inductive error** separates into three conceptually distinct sources:

$$
\begin{array}{rclll}
a_{\approx}(\mathcal{S}) - a(\mathcal{P}_{Target}) & = & [a_{\approx}(\mathcal{S}) - a(\mathcal{S})\ ] & \ldots & \text{measurement} \\
 & & +\ [a(\mathcal{S}) - a(\mathcal{P}_{Study})\ ] & \ldots & \text{sample}
\end{array}
$$

UNIVERSITY OF WATERLOO

Inductive inference

We draw conclusions about $a(\mathcal{P}_{Target})$ from an attribute $a_{\approx}(\mathcal{S}) = \widehat{a}(\mathcal{S})$, being based on measurements of units in some sample $\mathcal{S}$. A case needs to be made for the validity of each step along **the inductive path**:



The overall **inductive error** separates into three conceptually distinct sources:

$$
\begin{array}{rcllcl}
a_{\approx}(\mathcal{S}) - a(\mathcal{P}_{Target}) & = & [a_{\approx}(\mathcal{S}) - a(\mathcal{S})\ ] & & \ldots & \text{measurement} \\
 & & +\ [a(\mathcal{S}) - a(\mathcal{P}_{Study})\ ] & & \ldots & \text{sample} \\
 & & +\ [a(\mathcal{P}_{Study}) - a(\mathcal{P}_{Target})\ ] & & \ldots & \text{study}
\end{array}
$$

UNIVERSITY OF WATERLOO

Inductive error

$$
\begin{aligned}
a_{\approx}(\mathcal{S}) - a(\mathcal{P}_{Target}) \quad = \quad & [a_{\approx}(\mathcal{S}) - a(\mathcal{S})\ ] && \ldots \quad \text{measurement} \\
& +\ [a(\mathcal{S}) - a(\mathcal{P}_{Study})\ ] && \ldots \quad \text{sample} \\
& +\ [a(\mathcal{P}_{Study}) - a(\mathcal{P}_{Target})\ ] && \ldots \quad \text{study}
\end{aligned}
$$

Inductive error

$$
\begin{aligned}
a_{\approx}(\mathcal{S}) - a(\mathcal{P}_{Target}) \quad = \quad & [a_{\approx}(\mathcal{S}) - a(\mathcal{S})\ ] && \ldots \quad \text{measurement} \\
& +\ [a(\mathcal{S}) - a(\mathcal{P}_{Study})\ ] && \ldots \quad \text{sample} \\
& +\ [a(\mathcal{P}_{Study}) - a(\mathcal{P}_{Target})\ ] && \ldots \quad \text{study}
\end{aligned}
$$

- measurement error: reduced by increasing the accuracy of the measuring systems

Inductive error

$$
\begin{aligned}
a_{\approx}(\mathcal{S}) - a(\mathcal{P}_{Target}) \quad = \quad & [a_{\approx}(\mathcal{S}) - a(\mathcal{S})\ ] && \ldots \quad \text{measurement} \\
& +\ [a(\mathcal{S}) - a(\mathcal{P}_{Study})\ ] && \ldots \quad \text{sample} \\
& +\ [a(\mathcal{P}_{Study}) - a(\mathcal{P}_{Target})\ ] && \ldots \quad \text{study}
\end{aligned}
$$

- measurement error: reduced by increasing the accuracy of the measuring systems
- sample error: assurances from statistics
  - reduced sampling bias

UNIVERSITY OF WATERLOO

Inductive error

$$
\begin{array}{rcllcl}
a_{\approx}(\mathcal{S}) - a(\mathcal{P}_{Target}) & = & [a_{\approx}(\mathcal{S}) - a(\mathcal{S})\ ] & & \ldots & \text{measurement} \\
 & & +\ [a(\mathcal{S}) - a(\mathcal{P}_{Study})\ ] & & \ldots & \text{sample} \\
 & & +\ [a(\mathcal{P}_{Study}) - a(\mathcal{P}_{Target})\ ] & & \ldots & \text{study}
\end{array}
$$

- measurement error: reduced by increasing the accuracy of the measuring systems
- sample error: assurances from statistics
  - reduced sampling bias
  - reduced sampling variability

Inductive error

$$
\begin{aligned}
a_{\approx}(\mathcal{S}) - a(\mathcal{P}_{Target}) \quad = \quad & [a_{\approx}(\mathcal{S}) - a(\mathcal{S})\,] && \ldots \quad \text{measurement} \\
& +\ [a(\mathcal{S}) - a(\mathcal{P}_{Study})\,] && \ldots \quad \text{sample} \\
& +\ [a(\mathcal{P}_{Study}) - a(\mathcal{P}_{Target})\,] && \ldots \quad \text{study}
\end{aligned}
$$

- measurement error: reduced by increasing the accuracy of the measuring systems
- sample error: assurances from statistics
  - reduced sampling bias
  - reduced sampling variability
  - different sampling designs achieve these for various circumstances
  - these are more like insurance policies than guarantees
- study error: can be the most difficult case to make

UNIVERSITY OF WATERLOO

Inductive error

$$\begin{aligned}
a_{\approx}(\mathcal{S}) - a(\mathcal{P}_{Target}) \quad = \quad & [a_{\approx}(\mathcal{S}) - a(\mathcal{S})\,] && \ldots \quad \text{measurement} \\
& +\ [a(\mathcal{S}) - a(\mathcal{P}_{Study})\,] && \ldots \quad \text{sample} \\
& +\ [a(\mathcal{P}_{Study}) - a(\mathcal{P}_{Target})\,] && \ldots \quad \text{study}
\end{aligned}$$

- measurement error: reduced by increasing the accuracy of the measuring systems
- sample error: assurances from statistics
  - reduced sampling bias
  - reduced sampling variability
  - different sampling designs achieve these for various circumstances
  - these are more like insurance policies than guarantees
- study error: can be the most difficult case to make and can be statistical, non-statistical, or both.

UNIVERSITY OF WATERLOO

Example - Visible minorities in Canada 2006

Recall the `minority` data from loon.data.

Questions:

- What population attribute(s) are of interest?

Example - Visible minorities in Canada 2006

Recall the `minority` data from loon.data.

Questions:

- What population attribute(s) are of interest?
- What is the target population?

Example - Visible minorities in Canada 2006

Recall the `minority` data from loon.data.

Questions:

- What population attribute(s) are of interest?
- What is the target population?
- What is the study population?

Example - Visible minorities in Canada 2006

Recall the `minority` data from loon.data.

Questions:

- What population attribute(s) are of interest?
- What is the target population?
- What is the study population?
- What is the sample?

Example - Visible minorities in Canada 2006

Recall the `minority` data from loon.data.

Questions:

- What population attribute(s) are of interest?
- What is the target population?
- What is the study population?
- What is the sample?
- What is the sampling plan?

Example - Visible minorities in Canada 2006

Recall the `minority` data from loon.data.

Questions:

- What population attribute(s) are of interest?
- What is the target population?
- What is the study population?
- What is the sample?
- What is the sampling plan?

How these are answered will depend upon the definition of the units:

Example - Visible minorities in Canada 2006

Recall the `minority` data from loon.data.

Questions:

- What population attribute(s) are of interest?
- What is the target population?
- What is the study population?
- What is the sample?
- What is the sampling plan?

How these are answered will depend upon the definition of the units:

- a city is a unit?

Example - Visible minorities in Canada 2006

Recall the `minority` data from loon.data.

Questions:

- What population attribute(s) are of interest?
- What is the target population?
- What is the study population?
- What is the sample?
- What is the sampling plan?

How these are answered will depend upon the definition of the units:

- a city is a unit?
- person is a unit?

Example - Motor Trend cars 1974

Recall the `mtcars` data from R .

Questions:

- What population attribute(s) are of interest?

Example - Motor Trend cars 1974

Recall the `mtcars` data from R .

Questions:

- What population attribute(s) are of interest?
- What is the target population?

Example - Motor Trend cars 1974

Recall the `mtcars` data from R .

Questions:

- What population attribute(s) are of interest?
- What is the target population?
- What is the study population?

Example - Motor Trend cars 1974

Recall the `mtcars` data from R .

Questions:

- What population attribute(s) are of interest?
- What is the target population?
- What is the study population?
- What is the sample?

UNIVERSITY OF WATERLOO

Example - Motor Trend cars 1974

Recall the `mtcars` data from R .

Questions:

- What population attribute(s) are of interest?
- What is the target population?
- What is the study population?
- What is the sample?
- What is the sampling plan?

Example - Motor Trend cars 1974

Recall the `mtcars` data from R .

Questions:

- What population attribute(s) are of interest?
- What is the target population?
- What is the study population?
- What is the sample?
- What is the sampling plan?

How these are answered will depend upon the definition of the units:

Example - Motor Trend cars 1974

Recall the `mtcars` data from R .

Questions:

- What population attribute(s) are of interest?
- What is the target population?
- What is the study population?
- What is the sample?
- What is the sampling plan?

How these are answered will depend upon the definition of the units:

- a car is a unit?

Example - Human Immunoglobulin G1

Recall the `igg1` data from loon.data.

Questions:

- What population attribute(s) are of interest?

Example - Human Immunoglobulin G1

Recall the `igg1` data from loon.data.

Questions:

- What population attribute(s) are of interest?
- What is the target population?

Example - Human Immunoglobulin G1

Recall the `igg1` data from loon.data.

Questions:

- What population attribute(s) are of interest?
- What is the target population?
- What is the study population?

Example - Human Immunoglobulin G1

Recall the `igg1` data from loon.data.

Questions:

- What population attribute(s) are of interest?
- What is the target population?
- What is the study population?
- What is the sample?

Example - Human Immunoglobulin G1

Recall the `igg1` data from loon.data.

Questions:

- What population attribute(s) are of interest?
- What is the target population?
- What is the study population?
- What is the sample?
- What is the sampling plan?

Example - Human Immunoglobulin G1

Recall the `igg1` data from loon.data.

Questions:

- What population attribute(s) are of interest?
- What is the target population?
- What is the study population?
- What is the sample?
- What is the sampling plan?

This one is a little harder.

Example - November 3, 1936 US Presidential Election

Franklin D. Roosevelt (incumbent) versus Alf Landon

Example - November 3, 1936 US Presidential Election

Franklin D. Roosevelt (incumbent) versus Alf Landon The *Literary Digest* mailed 10 million questionnaires to known readers of *Literary Digest* and to potential readers (compiled via phonebooks, country club memberships, driver registrations).

Example - November 3, 1936 US Presidential Election

Franklin D. Roosevelt (incumbent) versus Alf Landon The *Literary Digest* mailed 10 million questionnaires to known readers of *Literary Digest* and to potential readers (compiled via phonebooks, country club memberships, driver registrations).

Of the 10 million sent out, 2.27 million questionnaires were answered and returned.

Example - November 3, 1936 US Presidential Election

Franklin D. Roosevelt (incumbent) versus Alf Landon The *Literary Digest* mailed 10 million questionnaires to known readers of *Literary Digest* and to potential readers (compiled via phonebooks, country club memberships, driver registrations).

Of the 10 million sent out, 2.27 million questionnaires were answered and returned.

Note that the *Literary Digest* had correctly predicted the winner for the previous 5 elections.

Example - November 3, 1936 US Presidential Election

Franklin D. Roosevelt (incumbent) versus Alf Landon The *Literary Digest* mailed 10 million questionnaires to known readers of *Literary Digest* and to potential readers (compiled via phonebooks, country club memberships, driver registrations).

Of the 10 million sent out, 2.27 million questionnaires were answered and returned.

Note that the *Literary Digest* had correctly predicted the winner for the previous 5 elections.

Questions:

- What are the units?

Example - November 3, 1936 US Presidential Election

Franklin D. Roosevelt (incumbent) versus Alf Landon The *Literary Digest* mailed 10 million questionnaires to known readers of *Literary Digest* and to potential readers (compiled via phonebooks, country club memberships, driver registrations).

Of the 10 million sent out, 2.27 million questionnaires were answered and returned.

Note that the *Literary Digest* had correctly predicted the winner for the previous 5 elections.

Questions:

- What are the units?
- What population attribute(s) are of interest?

Example - November 3, 1936 US Presidential Election

Franklin D. Roosevelt (incumbent) versus Alf Landon The *Literary Digest* mailed 10 million questionnaires to known readers of *Literary Digest* and to potential readers (compiled via phonebooks, country club memberships, driver registrations).

Of the 10 million sent out, 2.27 million questionnaires were answered and returned.

Note that the *Literary Digest* had correctly predicted the winner for the previous 5 elections.

Questions:

- What are the units?
- What population attribute(s) are of interest?
- What is the target population?

Example - November 3, 1936 US Presidential Election

Franklin D. Roosevelt (incumbent) versus Alf Landon The *Literary Digest* mailed 10 million questionnaires to known readers of *Literary Digest* and to potential readers (compiled via phonebooks, country club memberships, driver registrations).

Of the 10 million sent out, 2.27 million questionnaires were answered and returned.

Note that the *Literary Digest* had correctly predicted the winner for the previous 5 elections.

Questions:

- What are the units?
- What population attribute(s) are of interest?
- What is the target population?
- What is the study population?

Example - November 3, 1936 US Presidential Election

Franklin D. Roosevelt (incumbent) versus Alf Landon The *Literary Digest* mailed 10 million questionnaires to known readers of *Literary Digest* and to potential readers (compiled via phonebooks, country club memberships, driver registrations).

Of the 10 million sent out, 2.27 million questionnaires were answered and returned.

Note that the *Literary Digest* had correctly predicted the winner for the previous 5 elections.

Questions:

- What are the units?
- What population attribute(s) are of interest?
- What is the target population?
- What is the study population?
- What is the sample?

Example - November 3, 1936 US Presidential Election

Franklin D. Roosevelt (incumbent) versus Alf Landon The *Literary Digest* mailed 10 million questionnaires to known readers of *Literary Digest* and to potential readers (compiled via phonebooks, country club memberships, driver registrations).

Of the 10 million sent out, 2.27 million questionnaires were answered and returned.

Note that the *Literary Digest* had correctly predicted the winner for the previous 5 elections.

Questions:

- What are the units?
- What population attribute(s) are of interest?
- What is the target population?
- What is the study population?
- What is the sample?
- What is the sampling plan?

Example - November 3, 1936 US Presidential Election

Franklin D. Roosevelt (incumbent) versus Alf Landon The *Literary Digest* mailed 10 million questionnaires to known readers of *Literary Digest* and to potential readers (compiled via phonebooks, country club memberships, driver registrations).

Of the 10 million sent out, 2.27 million questionnaires were answered and returned.

Note that the *Literary Digest* had correctly predicted the winner for the previous 5 elections.

Questions:

- What are the units?
- What population attribute(s) are of interest?
- What is the target population?
- What is the study population?
- What is the sample?
- What is the sampling plan?

The October 31 issue of *Literary Digest* announced that Landon would be the winner with 57.1% of the vote and 370 electoral votes.

UNIVERSITY OF WATERLOO

Example - November 3, 1936 US Presidential Election

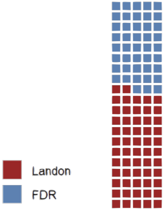
Results:

Poll Actual

Example - November 3, 1936 US Presidential Election

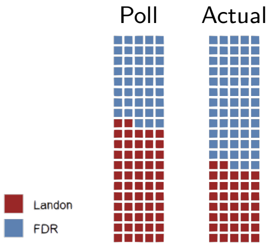
Results:

Poll Actual



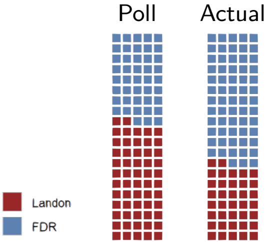
Example - November 3, 1936 US Presidential Election

Results:



Example - November 3, 1936 US Presidential Election

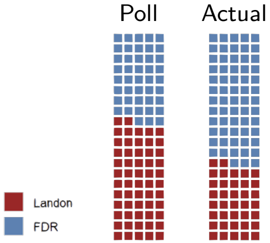
Results:



- Study Error?

Example - November 3, 1936 US Presidential Election

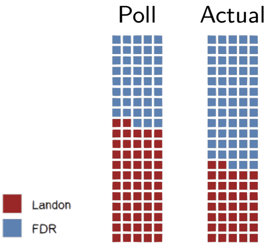
Results:



- Study Error? Possible problems?
- Sample Error?

Example - November 3, 1936 US Presidential Election

Results:

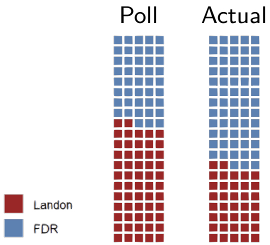


- Study Error? Possible problems?
- Sample Error? Possible problems?

UNIVERSITY OF WATERLOO

Example - November 3, 1936 US Presidential Election

Results:

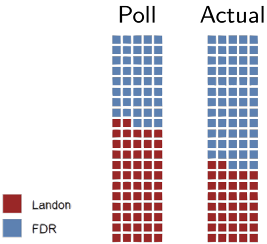


- Study Error? Possible problems?
- Sample Error? Possible problems?

Study error: Study population skewed towards wealthier voter.

Example - November 3, 1936 US Presidential Election

Results:



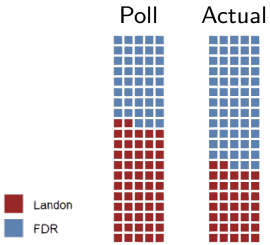
- Study Error? Possible problems?
- Sample Error? Possible problems?

Study error: Study population skewed towards wealthier voter.

Sample error: Sample is self-selected.

Example - November 3, 1936 US Presidential Election

Results:



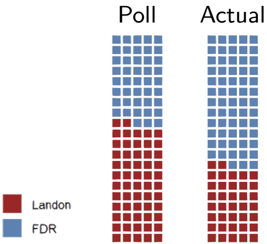
- Study Error? Possible problems?
- Sample Error? Possible problems?

Study error: Study population skewed towards wealthier voter.

Sample error: Sample is self-selected. Those who are most intensely interested will more likely respond.

Example - November 3, 1936 US Presidential Election

Results:



- Study Error? Possible problems?
- Sample Error? Possible problems?

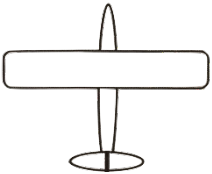
Study error: Study population skewed towards wealthier voter.

Sample error: Sample is self-selected. Those who are most intensely interested will more likely respond. "the minority of anti-Roosevelt voters felt more strongly than the pro-Roosevelt majority."

UNIVERSITY OF WATERLOO

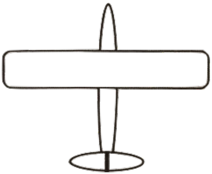
Example - World War II US Bombers

During the Second World War, US Statistician Abraham Wald was trying to determine where USAF bombers should have armour added to them in order to reduce the number of airplanes lost.



Example - World War II US Bombers

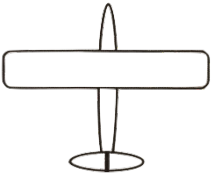
During the Second World War, US Statistician Abraham Wald was trying to determine where USAF bombers should have armour added to them in order to reduce the number of airplanes lost.



Trouble was that the only planes Wald had access to were those that returned from a mission, that were not shot down.

Example - World War II US Bombers

During the Second World War, US Statistician Abraham Wald was trying to determine where USAF bombers should have armour added to them in order to reduce the number of airplanes lost.

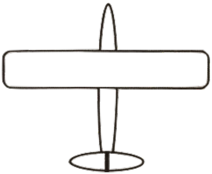


Trouble was that the only planes Wald had access to were those that returned from a mission, that were not shot down.

A **unit**, u, is an American bomber in the Second World War.

Example - World War II US Bombers

During the Second World War, US Statistician Abraham Wald was trying to determine where USAF bombers should have armour added to them in order to reduce the number of airplanes lost.



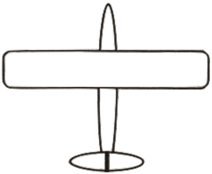
Trouble was that the only planes Wald had access to were those that returned from a mission, that were not shot down.

A **unit**, u, is an American bomber in the Second World War.

Variates are locations of bullet/flak holes on the plane.

Example - World War II US Bombers

During the Second World War, US Statistician Abraham Wald was trying to determine where USAF bombers should have armour added to them in order to reduce the number of airplanes lost.



Trouble was that the only planes Wald had access to were those that returned from a mission, that were not shot down.

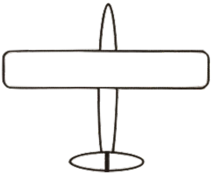
A **unit**, u, is an American bomber in the Second World War.

Variates are locations of bullet/flak holes on the plane.

Target population the planes that were shot down.

Example - World War II US Bombers

During the Second World War, US Statistician Abraham Wald was trying to determine where USAF bombers should have armour added to them in order to reduce the number of airplanes lost.



Trouble was that the only planes Wald had access to were those that returned from a mission, that were not shot down.

A **unit**, u, is an American bomber in the Second World War.

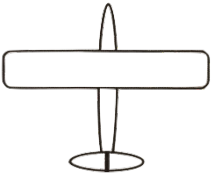
Variates are locations of bullet/flak holes on the plane.

Target population the planes that were shot down.

Study population the planes that actually return.

Example - World War II US Bombers

During the Second World War, US Statistician Abraham Wald was trying to determine where USAF bombers should have armour added to them in order to reduce the number of airplanes lost.



Trouble was that the only planes Wald had access to were those that returned from a mission, that were not shot down.

A **unit**, u, is an American bomber in the Second World War.

Variates are locations of bullet/flak holes on the plane.

Target population the planes that were shot down.

Study population the planes that actually return.

Sample returning planes that are accessible and measured.

UNIVERSITY OF WATERLOO

Example - World War II US Bombers

Wald's solution:

Example - World War II US Bombers

Wald's solution: For each returning plane, mark the locations of holes on a template.

Example - World War II US Bombers

Wald's solution: For each returning plane, mark the locations of holes on a template.

Combine all templates into a single graphical attribute:

Example - World War II US Bombers

Wald's solution: For each returning plane, mark the locations of holes on a template.

Combine all templates into a single graphical attribute:

Dark spots show holes on all returning planes in the sample.

Example - World War II US Bombers

Wald's solution: For each returning plane, mark the locations of holes on a template.

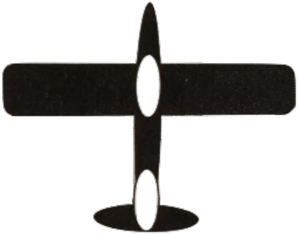
Combine all templates into a single graphical attribute:

Dark spots show holes on all returning planes in the sample. Clearly in error for the corresponding target population attribute.

Example - World War II US Bombers

Wald's solution: For each returning plane, mark the locations of holes on a template.

Combine all templates into a single graphical attribute:



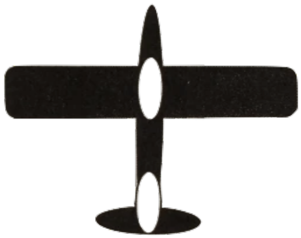
Dark spots show holes on all returning planes in the sample. Clearly in error for the corresponding target population attribute.

Because they didn't return, they must have holes where these did not.

Example - World War II US Bombers

Wald's solution: For each returning plane, mark the locations of holes on a template.

Combine all templates into a single graphical attribute:



Dark spots show holes on all returning planes in the sample. Clearly in error for the corresponding target population attribute.

Because they didn't return, they must have holes where these did not. Take advantage of known study error and add extra armour on these areas.

"Wald and his wife died when the Air India plane in which they were travelling crashed in the Nilgiri Mountains, in southern India, while on an extensive lecture tour at the invitation of the Indian government." – Wikipedia

Example - Kodak colour development 1954

Need to ensure that development process is properly calibrated.

Example - Kodak colour development 1954

Need to ensure that development process is properly calibrated.

Kodak supplies photos to studios to calibrate process.

Example - Kodak colour development 1954

Need to ensure that development process is properly calibrated.

Kodak supplies photos to studios to calibrate process.

Example - Kodak colour development 1954

Need to ensure that development process is properly calibrated.

Kodak supplies photos to studios to calibrate process.

Questions:

- What are the units?

Example - Kodak colour development 1954

Need to ensure that development process is properly calibrated.

Kodak supplies photos to studios to calibrate process.

Questions:

- What are the units?
- What population attribute(s) are of interest?

Example - Kodak colour development 1954

Need to ensure that development process is properly calibrated.

Kodak supplies photos to studios to calibrate process.

Questions:

- What are the units?
- What population attribute(s) are of interest?
- What is the target population?

Example - Kodak colour development 1954

Need to ensure that development process is properly calibrated.

Kodak supplies photos to studios to calibrate process.

Questions:

- What are the units?
- What population attribute(s) are of interest?
- What is the target population?
- What is the study population?

UNIVERSITY OF WATERLOO

Example - Kodak colour development 1954

Need to ensure that development process is properly calibrated.

Kodak supplies photos to studios to calibrate process.

Questions:

- What are the units?
- What population attribute(s) are of interest?
- What is the target population?
- What is the study population?
- What is the sample?

UNIVERSITY OF WATERLOO

Example - Kodak colour development 1954

Need to ensure that development process is properly calibrated.

Kodak supplies photos to studios to calibrate process.

Questions:

- What are the units?
- What population attribute(s) are of interest?
- What is the target population?
- What is the study population?
- What is the sample?
- What is the sampling plan?

UNIVERSITY OF WATERLOO

Example - Kodak colour development 1954

Often left darker tones looking washed out.

Example - Kodak colour development 1954

Often left darker tones looking washed out.

Example - Kodak colour development 1954

Often left darker tones looking washed out.

Example - Kodak colour development 1954

Often left darker tones looking washed out.

Sources of error?

Example - Crash testing cars

Crash test dummies have been used to test the effects of car crashes on occupants in the car since they were introduced by Chevrolet in the 1950s.

Example - Crash testing cars

Crash test dummies have been used to test the effects of car crashes on occupants in the car since they were introduced by Chevrolet in the 1950s.

In the 1970s, Chevrolet's Hybrid 2 dummy becomes the industry standard.

Example - Crash testing cars

Crash test dummies have been used to test the effects of car crashes on occupants in the car since they were introduced by Chevrolet in the 1950s.

In the 1970s, Chevrolet's Hybrid 2 dummy becomes the industry standard.

Modelled after the 50th percentile "average" adult male from USA population.

Example - Crash testing cars

Crash test dummies have been used to test the effects of car crashes on occupants in the car since they were introduced by Chevrolet in the 1950s.

In the 1970s, Chevrolet's Hybrid 2 dummy becomes the industry standard.

Modelled after the 50th percentile "average" adult male from USA population.

Hybrid II and Hybrid III

Example - Crash testing cars

Crash test dummies have been used to test the effects of car crashes on occupants in the car since they were introduced by Chevrolet in the 1950s.

In the 1970s, Chevrolet's Hybrid 2 dummy becomes the industry standard.

Modelled after the 50th percentile "average" adult male from USA population.

Hybrid II and Hybrid III

Questions:

- Units?

Example - Crash testing cars

Crash test dummies have been used to test the effects of car crashes on occupants in the car since they were introduced by Chevrolet in the 1950s.

In the 1970s, Chevrolet's Hybrid 2 dummy becomes the industry standard.

Modelled after the 50th percentile "average" adult male from USA population.

Hybrid II and Hybrid III

Questions:

- Units?
- Population attribute(s)?

UNIVERSITY OF WATERLOO

Example - Crash testing cars

Crash test dummies have been used to test the effects of car crashes on occupants in the car since they were introduced by Chevrolet in the 1950s.

In the 1970s, Chevrolet's Hybrid 2 dummy becomes the industry standard.

Modelled after the 50th percentile "average" adult male from USA population.

Hybrid II and Hybrid III

Questions:

- Units?
- Population attribute(s)?
- Target population?

Example - Crash testing cars

Crash test dummies have been used to test the effects of car crashes on occupants in the car since they were introduced by Chevrolet in the 1950s.

In the 1970s, Chevrolet's Hybrid 2 dummy becomes the industry standard.

Modelled after the 50th percentile "average" adult male from USA population.

Hybrid II and Hybrid III

Questions:

- Units?
- Population attribute(s)?
- Target population?
- Study population?

UNIVERSITY OF WATERLOO

Example - Crash testing cars

Crash test dummies have been used to test the effects of car crashes on occupants in the car since they were introduced by Chevrolet in the 1950s.

In the 1970s, Chevrolet's Hybrid 2 dummy becomes the industry standard.

Modelled after the 50th percentile "average" adult male from USA population.

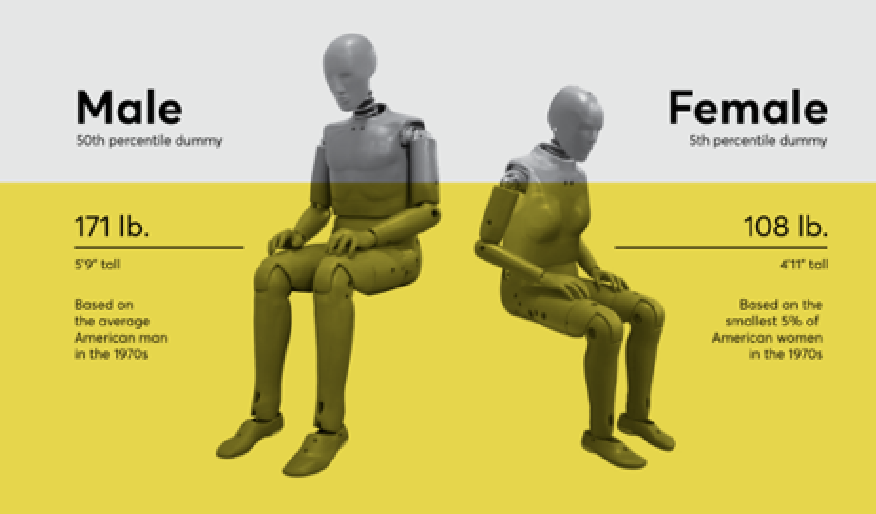
Hybrid II and Hybrid III

Questions:

- Units?
- Population attribute(s)?
- Target population?
- Study population?
- Sample?

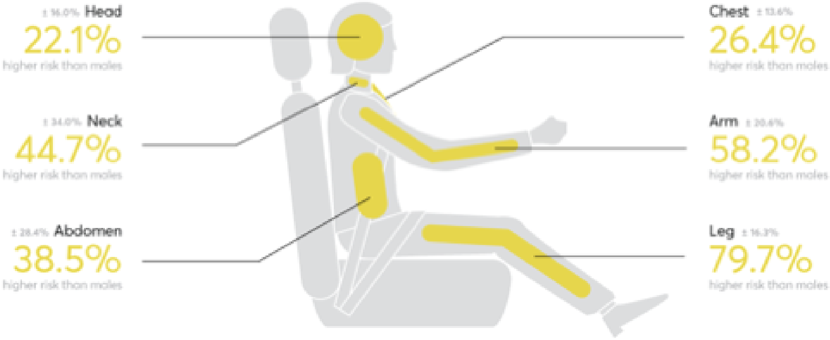
UNIVERSITY OF WATERLOO

Example - Problems



Male versus Female crash test dummmies

Example - Problems



Male versus Female crash tests

Examples - they go on and on

What are some more recent examples?

- Algorithms which filter job applicants

Examples - they go on and on

What are some more recent examples?

- Algorithms which filter job applicants
- Google's image tagging

Examples - they go on and on

What are some more recent examples?

- Algorithms which filter job applicants
- Google's image tagging
- Auto completion of search text

UNIVERSITY OF WATERLOO

Examples - they go on and on

What are some more recent examples?

- Algorithms which filter job applicants
- Google's image tagging
- Auto completion of search text
- Results returned by search

Examples - they go on and on

What are some more recent examples?

- Algorithms which filter job applicants
- Google's image tagging
- Auto completion of search text
- Results returned by search

For each of the above, what is the target population?

Examples - they go on and on

What are some more recent examples?

- Algorithms which filter job applicants
- Google's image tagging
- Auto completion of search text
- Results returned by search

For each of the above, what is the target population? The study population?

Examples - they go on and on

What are some more recent examples?

- Algorithms which filter job applicants
- Google's image tagging
- Auto completion of search text
- Results returned by search

For each of the above, what is the target population? The study population? The sample?

Examples - they go on and on

What are some more recent examples?

- Algorithms which filter job applicants
- Google's image tagging
- Auto completion of search text
- Results returned by search

For each of the above, what is the target population? The study population? The sample?

Are these the same for the user as for the provider?

Examples - they go on and on

What are some more recent examples?

- Algorithms which filter job applicants
- Google's image tagging
- Auto completion of search text
- Results returned by search

For each of the above, what is the target population? The study population? The sample?

Are these the same for the user as for the provider?

What problem do advocates of "Machine Learning Fairness" believe they are addressing?

Examples - they go on and on

What are some more recent examples?

- Algorithms which filter job applicants
- Google's image tagging
- Auto completion of search text
- Results returned by search

For each of the above, what is the target population? The study population? The sample?

Are these the same for the user as for the provider?

What problem do advocates of "Machine Learning Fairness" believe they are addressing?

Is "machine learning fairness" ethical?